

Современный
Гуманитарный
Университет

Дистанционное образование

0664.04.01; 1

Рабочий учебник

Фамилия, имя, отчество _____

Факультет _____

Номер контракта _____

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ЮНИТА 4

ПРИКЛАДНАЯ СТАТИСТИКА

Москва 2001

Разработано И.Б. Чернышевой, канд. техн. наук

Рекомендовано Министерством
образования Российской Федерации
в качестве учебного пособия для
студентов высших учебных заведений

КУРС: ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

- Юнита 1. Введение в теорию вероятностей.
- Юнита 2. Многомерные распределения и предельные теоремы.
- Юнита 3. Основные понятия математической статистики.
- Юнита 4. Прикладная статистика.

ЮНИТА 4

Рассматриваются основные понятия прикладной статистики: дисперсия распределения и ее эмпирическая оценка, проверка гипотез о законе распределения, дисперсионный анализ, корреляционный и регрессионный анализ, корреляционная модель.

Для студентов Современного Гуманитарного Университета

Юнита соответствует профессиональной образовательной программе № 1

ОГЛАВЛЕНИЕ

ДИДАКТИЧЕСКИЙ ПЛАН	5
ЛИТЕРАТУРА	6
ПЕРЕЧЕНЬ УМЕНИЙ	7
ТЕМАТИЧЕСКИЙ ОБЗОР	16
1. Изучение дисперсии распределения с помощью статистического эксперимента	16
1.1. Дисперсия и ее эмпирическая оценка	16
1.2. Два важных закона распределения	19
1.3. Распределение выборочной дисперсии	21
1.4. Интервальная оценка для дисперсии	22
1.5. Проверка гипотез, связанных с дисперсией	26
1.5.1. Гипотеза о числовом значении дисперсии	26
1.5.2. Сравнение дисперсий двух совокупностей	28
1.6. Оценка необходимого числа опытов для сравнения двух дисперсий	31
2. Проверка гипотез о законе распределения	31
2.1. Критерий согласия Пирсона (χ^2 -критерий)	32
2.2. Проверка гипотезы о том, что генеральное распределение нормальное	35
2.3. Проверка гипотезы о равномерном распределении	38
2.4. Критерий Колмогорова	39
2.5. Проверка гипотез об однородности выборок	41
2.6. Проверка гипотез об однородности выборок с помощью критерия χ^2	43
2.7. Критерий независимости	45
3. Дисперсионный анализ. Сравнение параметров более чем двух случайных величин	46
4. Корреляционный и регрессионный анализ. Выявление связи между величинами	59
4.1. Двумерные дискретные случайные величины. Понятие условного распределения. Условные математическое ожидание и дисперсия	59
4.2. Регрессия, линейная регрессия. Коэффициент корреляции	61
4.3. Двумерные непрерывные случайные величины	64
4.4. Описание выборок из двумерных случайных величин	66
4.4.1. Корреляционные таблицы	66
4.4.2. Выборочные условные средние	68

5. Элементы регрессионного анализа	69
5.1. Установление формы связи между переменными, прогнозы	69
5.2. Метод наименьших квадратов	70
5.3. Построение прямой методом наименьших квадратов ..	71
5.4. Проверка значимости уравнения регрессии	77
5.5. Нелинейная регрессия	78
6. Корреляционная модель	79
6.1. Две регрессионные кривые	79
6.2. Эмпирический коэффициент корреляции	82
6.3. Множественная корреляция	92
ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ	93
ТРЕНИНГ УМЕНИЙ	107
ФАЙЛ МАТЕРИАЛОВ	135
ГЛОССАРИЙ*	

ДИДАКТИЧЕСКИЙ ПЛАН

Дисперсия генеральная и выборочная. Распределения χ^2 и Фишера-Снедекора. Закон распределения выборочной дисперсии. Доверительные интервалы для дисперсии нормальной величины.

Проверка гипотез, связанных с дисперсией. Гипотеза о числовом значении дисперсии. Сравнение дисперсий двух совокупностей. Оценка необходимого числа опытов для сравнения двух дисперсий.

Проверка гипотез о законе распределения. Критерий согласия Пирсона (χ^2 -критерий). Критерий Колмогорова. Проверка гипотез об однородности выборок. Критерий независимости.

Дисперсионный анализ. Изучение влияния одного фактора. Сравнение параметров более чем двух случайных величин. Проверка гипотезы о равенстве m математических ожиданий. Проверка гипотезы о равенстве m дисперсий с помощью критерия χ^2 . Что изучает многофакторный анализ.

Корреляционный и регрессионный анализ. Регрессия, линейная регрессия. Коэффициент корреляции. Описание выборок из двумерных случайных величин. Корреляционные таблицы. Метод наименьших квадратов. Построение прямой регрессии методом наименьших квадратов. Прогнозы. Доверительные интервалы для прогнозируемого среднего и индивидуального значения. Проверка значимости уравнения регрессии. Эмпирический коэффициент корреляции как оценка генерального, интервальная оценка для генерального коэффициента корреляции.

ЛИТЕРАТУРА

Основная

- *1. Гмурман В.Е. Теория вероятностей и математическая статистика. М., 2000.
- *2. Кремер Н.Ш. Теория вероятностей и математическая статистика. М., 2000.

Дополнительная

- 3. Калинина В.Н., Панкин В.Ф. Математическая статистика. М., 1998.
- 4. Елисеева И.И., Юзбашев М.М. Общая теория статистики. М., 1996.
- 5. Колемаев В.А., Староверов О.В., Турундаевский В.Б. Теория вероятностей и математическая статистика. М., 1991.
- 6. Янко Я. Математико-статистические таблицы. М., 1961.

Задачники

- 7. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. М., 1997.
- 8. Булдык Г.М., Мацкевич И.П., Свирид Г.П. Сборник задач и упражнений по высшей математике (теория вероятностей и математическая статистика). Минск, 1996.

ПЕРЕЧЕНЬ УМЕНИЙ

№ п/п	Умение	Алгоритмы
1.	Вычисление по выборке точечной оценки для генеральной дисперсии σ^2 и среднеквадратического отклонения и доверительных интервалов для них. Проверка гипотезы о числовом значении дисперсии	1. Вычислить точечную оценку для среднего по формулам: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ – для выборки, заданной вариационным рядом; $\bar{x} = \frac{1}{n} \sum_{j=1}^k x_j m_j = \sum_{j=1}^k x_j \frac{m_j}{n} = \sum_{j=1}^k x_j \tilde{p}_j$ – для выборки, заданной таблицей. 2. Вычислить смещенную точечную оценку для дисперсии по формулам: $S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$ – для выборки, заданной вариационным рядом; $S^2 = \frac{1}{n} \sum_{j=1}^k x_j^2 m_j - \bar{x}^2 = \sum_{j=1}^k x_j^2 \frac{m_j}{n} - \bar{x}^2 = \sum_{j=1}^k (x_j - \bar{x})^2 \frac{m_j}{n}$ – для выборки, заданной таблицей 3. Вычислить несмещенную точечную оценку для дисперсии по формуле $s^2 = \frac{n}{n-1} S^2; \sigma = \sqrt{s^2}$ – оценка для среднеквадратического отклонения. 4. Вычислить границы доверительного интервала для генеральной дисперсии, используя таблицу 1 приложения 2. 5. Если требуется двусторонняя проверка гипотезы о числовом значении дисперсии (альтернативная гипотеза $H_1: \sigma^2 = \sigma_1^2 \neq \sigma_0^2$), то ее можно осуществить с помощью построенного доверительного интервала для дисперсий. В общем случае можно воспользоваться таблицами критических значений распределения χ^2. Критерии проверки для гипотезы H_0, состоящей в том, что $\sigma = \sigma_0$, по статистике $\chi^2 = \frac{nS^2}{\sigma_0^2}$ для уровня значимости α таковы:

№ п/п	Умение	Алгоритмы
		$H_1: \sigma^2 = \sigma_1^2 > \sigma_0^2 - \chi^2 > \chi_{\alpha, n-1}^2$; $H_1: \sigma^2 = \sigma_1^2 < \sigma_0^2 - \chi^2 > \chi_{1-\alpha, n-1}^2$; $H_1: \sigma^2 = \sigma_1^2 \neq \sigma_0^2 - \chi^2 > \chi_{\alpha/2, n-1}^2$ либо $\chi^2 < \chi_{1-\alpha/2, n-1}^2$. Граница критической области $\chi_{\alpha, k}^2$ берется из таблицы критических точек распределения χ^2 (таблица 2 приложения 2)
2.	Проверка статистических гипотез о совпадении двух или нескольких дисперсий с помощью выборок из нескольких генеральных совокупностей	Проверка гипотез о совпадении дисперсий двух совокупностей 1. Вычислить несмещенные оценки для генеральных дисперсий обеих совокупностей s_1^2 и s_2^2 (см. умение №1) и статистику для проверки гипотезы – их отношение $F = s_1^2/s_2^2$ 2. Выполнить проверку гипотезы. Критерии проверки для гипотезы H_0 , состоящей в том, что $\sigma_1^2 = \sigma_2^2$, по статистике $F = \frac{s_1^2}{s_2^2}$ для уровня значимости α таковы: $H_1: \sigma^2 = \sigma_1^2 > \sigma_2^2 - F > F_{\alpha, n_1-1, n_2-1}$; $H_1: \sigma^2 = \sigma_1^2 \neq \sigma_2^2 - F > F_{\alpha/2, n_1-1, n_2-1}$ либо $F < F_{1-\alpha/2, n_1-1, n_2-1}$ $\left(F < \frac{1}{F_{\alpha/2, n_2-1, n_1-1}} \right)$. Граница критической области F_{α, n_1-1, n_2-1} берется из таблицы критических точек распределения F (таблица 3 приложения 2). Для отыскания значений $F_{1-\alpha/2, n_1-1, n_2-1}$ надо воспользоваться равенством для $P > 0,5$ $F_p(k_1, k_2) = \frac{1}{F_{1-p}(k_2, k_1)}$. (очевидным, если выборки поменять местами). Проверка нулевой гипотезы о равенстве дисперсий m нормально распределенных совокупностей может быть выполнена с помощью критерия Бартлетта (см. главу 3 «Дисперсионный анализ»)

№ п/п	Умение	Алгоритмы
3.	Проверка статистических гипотез о законе распределения с помощью критерия согласия Пирсона или критерия Колмогорова. Проверка статистических гипотез об однородности двух выборок с помощью критерия Колмогорова-Смирнова и гипотезы об однородности m выборок с помощью критерия χ^2	1. Критерий согласия Пирсона (χ^2 – критерий). 1.1. Вычислить значение статистики U $U = \chi^2 = \sum_{i=1}^m \frac{(n_i - np_i)^2}{np_i},$ где n – число испытаний. В дискретном случае генеральные частоты p_i – это вероятности значений x_i и n_i частоты (сколько раз значение x_i встретилось в выборке). Если генеральное распределение непрерывное, то выборка группируется. Для интервалов группировки выборки (x_i, x_{i+1}) подсчитываются частоты n_i (число попаданий данных опыта в интервал) и теоретические вероятности p_i попадания в интервал (x_i, x_{i+1}). Для вычисления вероятностей p_i надо воспользоваться таблицами генерального распределения $F(x)$. $p_i = F(x_{i+1}) - F(x_i)$. 1.2. Для выбранного уровня значимости α по таблице χ^2 распределения (таблица 2 приложения 2) находят критическое значение $\chi^2_{\alpha, k}$ для числа степеней свободы $k = m - r - 1$ (m – число интервалов группировки или принимаемых значений, r – число параметров генерального распределения, замененных оценками). Сравнивают U и $\chi^2_{\alpha, k}$. Если $U > \chi^2_{\alpha, k}$, то гипотеза H_0 отвергается; если $U < \chi^2_{\alpha, k}$, гипотеза H_0 не противоречит опытным данным. 2. Критерий Колмогорова. В качестве меры расхождения между теоретическим и эмпирическим распределениями рассматривают максимальное значение абсолютной величины разности между эмпирической функцией распределения $F_n(x)$ и соответствующей теоретической функцией распределения $F(x)$ $D = \max F_n(x) - F(x)$, которое имеет распределение Колмогорова. Взяв, например, $t=1,36$, из таблиц функции Колмогорова (таблица 10 приложения 2) получаем, что при достаточно большом объеме выборки, если гипотеза о генеральном распределении верна, неравенство

№ п/п	Умение	Алгоритмы
		$F_n(x) - F(x) < \frac{1,36}{\sqrt{n}}$ (или $\sqrt{n} F_n(x) - F(x) < 1,36$) выполняется с вероятностью 0,95 (т.е. в среднем для 95% выборок). Следовательно, если указанное неравенство выполняется, гипотеза H_0 не противоречит данным опыта на уровне значимости 0,05. В противном случае она должна быть отклонена. 3. Проверка однородности двух выборок по критерию Колмогорова-Смирнова. Сравнить две эмпирические функции распределения. Эмпирическая функция распределения – это накопленные частоты, деленные на число испытаний, т.е. накопленные относительные частоты. Статистика критерия Колмогорова-Смирнова имеет вид: $\lambda' = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot \max F_{n_1}(x) - F_{n_2}(x) ,$ где $F_{n_1}(x)$ и $F_{n_2}(x)$ – эмпирические функции распределения, построенные по двум выборкам объемов n_1 и n_2. Гипотеза H_0 отвергается, если фактически наблюдаемое значение статистики λ' больше $\lambda'_{кр}$, и принимается в противном случае. При $n_1 > 50$ и $n_2 > 50$ статистика λ' имеет практически распределение Колмогорова. При меньших n надо пользоваться специальными таблицами. 4. Проверка однородности k выборок по критерию χ^2. Для проверки гипотезы об однородности k выборок надо вычислить значение статистики: $\chi_n^2 = \sum_{j=1}^k \sum_{i=1}^m \frac{(n_{ij} - n_j \cdot p_i)^2}{n_j \cdot p_i} = n \sum_{j=1}^k \sum_{i=1}^m \frac{\left(n_{ij} - 1/n \cdot \left(\sum_{j=1}^k n_{ij} \cdot \sum_{i=1}^m n_{ij} \right) \right)^2}{\sum_{j=1}^k n_{ij} \cdot \sum_{i=1}^m n_{ij}},$ где $p_i = \frac{1}{n} \sum_{j=1}^k n_{ij}$; $n = \sum_{j=1}^k \sum_{i=1}^m n_{ij}$ и воспользоваться таблицей критических точек распределения χ_n^2 для числа степеней свободы $n = (m-1)(k-1).$

№ п/п	Умение	Алгоритмы
4.	Проверка гипотезы о равенстве m математических ожиданий (элементы дисперсионного анализа)	Проверка выполняется средствами дисперсионного анализа. В дисперсионном анализе для каждого из j ($j=1,2,\dots,m$) уровней фактора проводится n испытаний (для простоты будем считать, что для всех уровней число опытов одно и то же и равно n). Пусть x_{ij} – значение измеряемого признака X, полученное в i-м опыте при j-м уровне фактора. Можно перейти от x_{ij} к $y_{ij}=10^k x_{ij}-C$. Результаты наблюдений располагаем в двумерной таблице размерности $n \times m$. Будем считать, что данные заданы так, что i (номер строки) – это номер наблюдения, а j (номер столбца) – это номер уровня фактора. 1. Для каждого уровня фактора j (столбца) вычислить суммированием по столбцу: $R_j = \sum_{i=1}^n x_{ij}^2; T_j = \sum_{i=1}^n x_{ij} \text{ или } R_j = \sum_{i=1}^n y_{ij}^2; T_j = \sum_{i=1}^n y_{ij}, \text{ если мы}$ перешли от x_{ij} к $y_{ij}=10^k x_{ij}-C$. 2. Выполнить суммирование по строке трех последних строк. 3. Вычислить Q, Q_1, Q_2, используя результаты суммирования по строке последних трех строк: $Q = \sum_{j=1}^m \sum_{i=1}^n x_{ij}^2 - \frac{1}{mn} \left(\sum_{j=1}^m \sum_{i=1}^n x_{ij} \right)^2 = \sum_{j=1}^m R_j - \frac{1}{mn} \left(\sum_{j=1}^m T_j \right)^2;$ $Q_1 = \frac{1}{n} \sum_{j=1}^m \left(\sum_{i=1}^n x_{ij} \right)^2 - \frac{1}{mn} \left(\sum_{j=1}^m \sum_{i=1}^n x_{ij} \right)^2 = \frac{1}{n} \sum_{j=1}^m T_j^2 - \frac{1}{mn} \left(\sum_{j=1}^m T_j \right)^2;$ $Q_{\text{ост}} = Q_{\text{общ}} - Q_{\text{факт}}$ или $Q_2 = Q - Q_1$. 4. Найти несмещенные оценки факторной и остаточной дисперсий. Для этого разделить Q_1 и Q_2 на соответствующее число степеней свободы. Число степеней свободы – это число квадратов в сумме минус число связывающих их уравнений. Поэтому для $s_{\text{факт}}^2 = s_1^2$ число степеней свободы равно $m-1$, а для $s_{\text{ост}}^2 = s_2^2 - mn - m$. Таким образом, $s_1^2 = Q_1/(m-1)$, а $s_2^2 = Q_2/(mn-m)$. 5. Вычислить значение статистики $F = \frac{s_1^2}{s_2^2}$. В таблице 3 приложения 2 найти критическое значение F_{α, k_1, k_2}.

№ п/п	Умение	Алгоритмы
		6. Принять решение о непротиворечивости гипотезы H_0. Гипотеза H_0 отвергается, если фактически вычисленное значение статистики $F > F_{\alpha; k_1; k_2}$, определенного на уровне значимости α при числе степеней свободы $k_1 = m - 1$, $k_2 = mn - m$, и принимается, если $F < F_{\alpha; k_1; k_2}$. Опровержение гипотезы означает наличие существенных различий в зависимости от уровня фактора. 7. Вычислить оценки для генерального среднего, если H_0 проходит, или генеральных средних, если H_0 не проходит. Если гипотеза H_0 о равенстве математических ожиданий при разных уровнях фактора не отвергается, т.е. влияние фактора A не подтверждается, то эффекты влияния уровней фактора λ_j будут нулевыми. Остальные параметры по результатам наблюдений оцениваются следующим образом: – оценка генерального среднего α равна общему среднему по всем наблюдениям $\bar{x} = \frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n x_{ij} = \frac{1}{mn} \sum_{j=1}^m T_j \quad (\text{в случае перехода к переменным } y_{ij} = 10^k x_{ij} - C \text{ следует не забыть сделать соответствующее преобразование, например, если } y_{ij} = x_{ij} - C, \text{ к этому значению надо добавить } C, \text{ т.е. в этом случае } \bar{x} = \frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n (y_{ij} + C) = C + \frac{1}{mn} \sum_{j=1}^m T_j);$ – оценка для σ^2: $s_{\text{ост}}^2 = \frac{1}{mn - m} \sum_{j=1}^m \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2.$ Если же гипотеза H_0 отвергается, т.е. влияние фактора не подтверждается по результатам наблюдений, то оценки для λ_j уже будут не нулевыми, а равными разности $\bar{x}_j - \bar{x}$, где $\bar{x}_j$ – групповое среднее на j-м уровне фактора A. 8. Вычислить коэффициент детерминации – отношение факторной вариации к общей (коэффициент детерминации указывает, какая доля дисперсии признака X объясняется зависимостью признака X от фактора): $r^2 = Q_{\text{факт}} / Q_{\text{общ.}}$ Выписать в словесной форме, что доля “r^2 дисперсии признака X объясняется зависимостью от фактора”.

№ п/п	Умение	Алгоритмы
5.	Вычисление коэффициента корреляции, графическое изображение функции взаимной зависимости между величинами с помощью прямой, построенной методом наименьших квадратов. Проверка значимости уравнения регрессии. [Выборочные значения пар (X,Y) заданы таблицей из двух строк].	1. Составить таблицу для расчета. В таблице должны присутствовать столбцы: Δx_i, $y_i \Delta x_i$, $(\Delta x_i)^2$, y_i^2, $ax_i + b$, $\Delta_i = y_i - ax_i - b$, $(\Delta_i)^2$ и две последние строки – для сумм по столбцу и этих сумм, деленных на n. 2. Вычислить коэффициенты a и b МНК-прямой. Уравнение прямой ищем в виде: $y = a(x - \bar{x}) + \bar{y},$ где $a = \frac{\frac{1}{n} \sum x_i y_i - \frac{1}{n} \sum x_i \cdot \frac{1}{n} \sum y_i}{\frac{1}{n} \sum x_i^2 - \left(\frac{1}{n} \sum x_i \right)^2} = \frac{\sum y_i \Delta x_i}{\sum (\Delta x_i)^2};$ $\Delta x_i = x_i - \bar{x}; \quad b = \bar{y} - a \cdot \bar{x}; \quad \bar{x} = \frac{1}{n} \sum x_i; \quad \bar{y} = \frac{1}{n} \sum y_i.$ 3. Построить график МНК-прямой (проводить прямую через две точки $(\bar{x}, \bar{y})$ и $(x_i, ax_i + b)$ с любым x_i). 4. Вычислить коэффициент корреляции r. Для системы точек, заданной таблицей из двух строк, он вычисляется по формуле $r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} =$ $= \frac{\frac{1}{n} \sum_{i=1}^n (x_i y_i) - \bar{x} \cdot \bar{y}}{\sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{y}^2}} = \frac{1}{n} \sum y_i \Delta x_i}{S_x S_y}.$ 5. Вычислить прогнозируемое значение, его доверительный интервал. Доверительный интервал для прогнозируемого значения вычисляется по формуле $ax + b - t_{1-\alpha; n-2} \cdot S_{Y(x)} < Y(x) < ax + b + t_{1-\alpha; n-2} \cdot S_{Y(x)},$ где

№ п/п	Умение	Алгоритмы
		$s^2_{Y(x)} = S^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x - \bar{x})^2} \right]; -$ $S^2 = 1/(n-2) \sum_{i=1}^n (ax_i + b - y_i)^2$ – оценка дисперсии σ^2. 6. Проверить значимость уравнения регрессии. Проверка значимости уравнения регрессии: $Q_R = \sum_{i=1}^n (ax_i + b - \bar{y})^2$ - сумма квадратов отклонений, обусловленная регрессией; $Q_{ост} = \sum_{i=1}^n [y_i - (ax_i + b)]^2$ - остаточная сумма квадратов, характеризующая отклонение от среднего в точке x_i, т.е. влияние неучтенных факторов. Уравнение регрессии значимо на уровне α, если фактически наблюдаемое значение статистики $F = \frac{Q_R (n - m)}{Q_{ост} (m - 1)} = \frac{Q_R (n - 2)}{Q_{ост} (2 - 1)} > F_{\alpha; 1; n-2},$ где $F_{\alpha; 1; n-2}$ – табличное значение F-критерия Фишера-Снедекора, определенное на уровне значимости α; n – число наблюдений. Число m оцениваемых по выборке параметров в нашем случае равно двум (a и b). Значение F показывает, в какой мере регрессия лучше оценивает значение зависимой переменной по сравнению со средней (именно поэтому в данном случае проверяемая гипотеза проходит при попадании статистики F в критическую область).
6.	Вычисление коэффициента корреляции, построение двух прямых регрессии. Выводы о	1. Вычислить (если они не заданы) в корреляционной таблице для каждого x_i значения $n_{i\cdot} = \sum_{j=1}^m n_{ij}$ (сколько раз такое значение x_i встретилось в паре с любыми y_j) и для каждого y_j значения $n_{\cdot j} = \sum_{i=1}^k n_{ij}$ (если X заданы по вертикали, а Y – по горизонтали; если же X заданы

№ п/п	Умение	Алгоритмы
	тесноте связи между компонентами. [Выборочные значения пар (X,Y) заданы корреляционной таблицей из m строк и k столбцов].	по горизонтали, а Y – по вертикали, то индексы просто меняются местами). 2. Вычислить: $\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_{i\cdot}, \quad s_x^2 = \frac{\sum_{i=1}^k x_i^2 n_{i\cdot}}{n} - (\bar{x})^2; \quad s_x = \sqrt{s_x^2};$ $\bar{y} = \frac{1}{n} \sum_{j=1}^m y_j n_{\cdot j}, \quad s_y^2 = \frac{\sum_{j=1}^m y_j^2 n_{\cdot j}}{n} - (\bar{y})^2; \quad s_y = \sqrt{s_y^2};$ $\mu = \overline{xy} - \bar{x} \cdot \bar{y} = \frac{\sum_{i=1}^k \sum_{j=1}^m x_i y_j n_{ij}}{n} - \bar{x} \cdot \bar{y};$ $r_{xy} = \frac{\mu}{s_x \cdot s_y}.$ Значимость коэффициента корреляции (то, что генеральный коэффициент корреляции $\rho \neq 0$) проверяется с помощью статистики: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. Гипотеза о том, что $\rho=0$, отвергается; если $ t = \frac{ r \sqrt{n-2}}{\sqrt{1-r^2}} > t_{1-\alpha; n-2}, \text{ где } t_{1-\alpha; n-2} - \text{табличное значение } t\text{-критерия Стьюдента, определенное на уровне значимости } \alpha \text{ при числе степеней свободы } n-2.$ 3. Выписать уравнения регрессии через коэффициенты a_{yx} и a_{xy} прямых регрессии, выписать тангенсы угла наклона прямых регрессии по формулам: $y_x - \bar{y} = a_{yx}(x - \bar{x}) = \frac{\mu}{s_x^2}(x - \bar{x}); \quad k_1 = a_{yx};$ $x_y - \bar{x} = a_{xy}(y - \bar{y}) = \frac{\mu}{s_y^2}(y - \bar{y}); \quad k_2 = \frac{1}{a_{xy}}.$ 4. Построить графики прямых регрессии [проводить прямые через точку $(\bar{x}, \bar{y})$ и еще одну точку, лежащую на прямой регрессии, как это делалось при построении МНК-прямой]. Сформулировать вывод о тесноте связи между компонентами.

1. ИЗУЧЕНИЕ ДИСПЕРСИИ РАСПРЕДЕЛЕНИЯ С ПОМОЩЬЮ СТАТИСТИЧЕСКОГО ЭКСПЕРИМЕНТА

1.1. Дисперсия и ее эмпирическая оценка

В предыдущей юните мы занимались некоторыми типичными задачами построения оценок для параметров генеральной совокупности по данным выборки и изучали их точность и надежность. То есть мы считали, что известен вид теоретического (генерального) распределения, но не известны и подлежат оценке параметры этого распределения. Так, например, для биномиального распределения по результатам n раз проведенного эксперимента мы оценивали значение p . Следующая задача состояла в том, что по результатам испытаний строились оценки для математического ожидания в случае, когда n раз проводятся испытания над величиной, которая распределена нормально. Для описания точности и надежности оценок мы пользовались тем, что оценка имеет нормальное распределение или, в случае малых выборок и неизвестной генеральной дисперсии, распределение Стьюдента. Мы не касались вопроса об оценке для дисперсии. Вопрос об оценке дисперсии очень важен. Вспомним, что нормальное распределение целиком определяется своими двумя параметрами – математическим ожиданием и дисперсией (или среднеквадратическим отклонением). Дисперсия характеризует такие важные показатели, как точность приборов, технологических процессов, степень однородности, риск, связанный с отклонением доходности активов от ожидаемого уровня и т.д.

В силу важности математического ожидания и дисперсии напомним их определение и важнейшие свойства (табл. 1.1, 1.2), которыми мы будем в дальнейшем пользоваться в наших рассуждениях. Поскольку размерность дисперсии случайной величины равна квадрату размерности самой случайной величины, чтобы охарактеризовать разброс, рассеяние случайной величины, чаще всего применяют даже не дисперсию $D(\xi)$, а **среднеквадратическое (стандартное) отклонение**

$\sigma = \sqrt{D(\xi)}$. Эта характеристика имеет ту же размерность, что и сама случайная величина.

* Жирным шрифтом выделены новые понятия, которые необходимо усвоить. Знание этих понятий будет проверяться при тестировании.

Таблица 1.1

Определение математического ожидания и дисперсии

	Дискретное распределение	Непрерывное распределение
Математическое ожидание $M(\xi) = a$	$\sum_{i=1}^n x_i \cdot p_i$	$\int_{-\infty}^{\infty} x \cdot f(x) dx$
Дисперсия $D(\xi) = \sigma^2$	$\begin{aligned} \sum_{i=1}^n (x_i - a)^2 \cdot p_i = \\ = \sum_{i=1}^n x_i^2 \cdot p_i - a^2 \end{aligned}$	$\begin{aligned} \int_{-\infty}^{\infty} (x - a)^2 \cdot f(x) dx = \\ = \int_{-\infty}^{\infty} x^2 \cdot f(x) dx - a^2 \end{aligned}$

Таблица 1.2

Свойства математического ожидания и дисперсии

	Математическое ожидание	Дисперсия	Среднеквадратиче- ское отклонение $\sigma = \sqrt{D(\xi)}$
Случайная величина (константа)	$M(C)=C$	$D(C)=0$	$\sigma(C)=0$
Случайная величина умножается на константу	$M(C \cdot \xi) = C \cdot M(\xi)$	$D(C \cdot \xi) = C^2 \cdot D(\xi)$	$\sigma(C \cdot \xi) = C \sigma(\xi)$
Случайные величины складываются	$M\left(\sum_{i=1}^n \xi_i\right) = \sum_{i=1}^n M(\xi_i)$	$D\left(\sum_{i=1}^n \xi_i\right) = \sum_{i=1}^n D(\xi_i)$ (для независимых слагаемых)	$\begin{aligned} \sigma(\xi_1 \pm \xi_2) = \\ = \sqrt{\sigma^2(\xi_1) + \sigma^2(\xi_2)} \end{aligned}$ (для независимых слагаемых)

Вычисление эмпирической дисперсии.

Выборка может быть задана последовательностью полученных значений $x_1, x_2, \dots, x_n$, либо таблицей частот (табл. 1.3), либо таблицей относительных частот, т.е. таблицей эмпирического распределения выборки (табл. 1.4).

Таблица 1.3

Значения x_i	x_1	x_2	...	x_k
Частоты m_i	m_1	m_2	...	m_k

Таблица 1.4

Значения x_i	x_1	x_2	...	x_k
Относительные частоты m_i/n	m_1/n	m_2/n	...	m_k/n

Формулы, по которым в зависимости от описания данных выборки вычисляются среднее значение и разброс выборки, приведены в табл. 1.5.

Таблица 1.5

	Вариационный ряд задан последовательно-стью	Задана таблица частот вариационного ряда	Задана таблица относительных частот вариационного ряда
Среднее значение выборки $\bar{x}$	$\frac{1}{n} \sum_{i=1}^n x_i$	$\frac{1}{n} \sum_{j=1}^k x_j m_j$	$\sum_{j=1}^k x_j \frac{m_j}{n}$
Дисперсия (разброс) выборки S^2	$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$	$\frac{1}{n} \sum_{j=1}^k x_j^2 m_j - \bar{x}^2 = \frac{1}{n} \sum_{j=1}^k (x_j - \bar{x})^2 m_j$	$\sum_{j=1}^k x_j^2 \frac{m_j}{n} - \bar{x}^2 = \sum_{j=1}^k (x_j - \bar{x})^2 \frac{m_j}{n}$

Пример 1.1. По выборке 4, 6, 7, 7, 10, 15, 18 ($n=7$) найти $\bar{x}$ и S^2 .

$$\bar{x} = (4+6+7+7+10+15+18)/7=9,57;$$

$$S^2 = 1/7(16+36+49+49+100+225+324)-(9,57)^2 = 114,14 - 91,58 = 22,56$$

Пример 1.2. Найти $\bar{x}$ и S^2 по таблице выборки:

Варианты	2	6	12	
m_i	3	10	7	$n=20$

$$\bar{x} = (2 \cdot 3 + 6 \cdot 10 + 12 \cdot 7) / 20 = 7,5;$$

$$S^2 = 1/20(43 + 36 \cdot 10 + 144 \cdot 7) - (7,5)^2 = (1/20) \cdot 1380 - 56,25 = 69 - 56,25 = 12,75$$

Пример 1.3. Найти $\bar{x}$ и S^2 по таблице выборки:

Варианты	2	6	12	
m_i/n	0,15	0,5	0,35	1

$$\bar{x} = 2 \cdot 0,15 + 6 \cdot 0,5 + 12 \cdot 0,35 = 7,5;$$

$$S^2 = (4 \cdot 0,15 + 36 \cdot 0,5 + 144 \cdot 0,35) - (7,5)^2 = 69 - 56,25 = 12,75$$

(легко видеть, что примеры 1.2 и 1.3 задают одну и ту же выборку, но в примере 1.2 она задана таблицей частот, а в примере 1.3 – таблицей относительных частот: $0,15=3/20$; $0,5=10/20$; $0,35=7/20$).

1.2. Два важных закона распределения

Нам понадобятся два распределения, важных для изучения свойств эмпирической дисперсии.

Распределение χ_n^2 (хи-квадрат) Пирсона – это распределение случайной величины, равной

$$\chi_n^2 = \xi_1^2 + \dots + \xi_n^2,$$

где ξ_i – независимые, нормально распределённые, стандартные $N(0,1)$ случайные величины. Число слагаемых n называется **числом степеней свободы** этого распределения.

Случайная величина, имеющая распределение χ_n^2 , принимает только положительные значения (рис. 1.1).

Это распределение хорошо изучено, и для него существуют подробные таблицы. При $n > 30$ распределение случайной величины

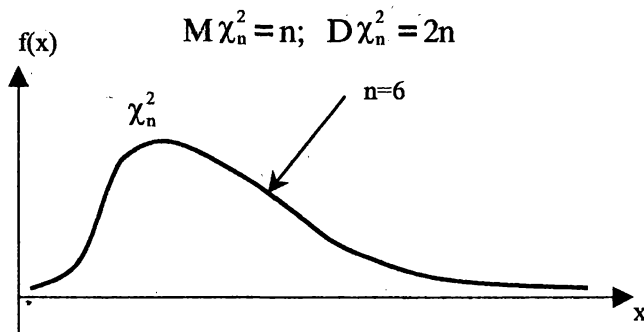


Рис. 1.1. График $f(x)$ – плотности χ_n^2 -распределения для $n = 6$

$\eta = \sqrt{2\chi_n^2} - \sqrt{2n-1}$ близко к стандартному нормальному $N(0,1)$. Поэтому обычно таблицы задаются для $n < 30$. Число степеней свободы определяют как разность между числом суммируемых случайных величин и числом линейных связей, ограничивающих свободу изменения этих величин. В сумме $\chi_n^2 = \xi_1^2 + \dots + \xi_n^2$ слагаемые независимы, число степеней свободы равно числу слагаемых, то есть n .

В таблице 2 приложения 2 содержатся значения функции распределения χ_n^2 для самых распространенных значений аргумента.

Распределением Фишера-Снедекора (или F-распределением) называется распределение случайной величины

$$F = \frac{\frac{1}{n_1} \chi_{n_1}^2}{\frac{1}{n_2} \chi_{n_2}^2},$$

где $\chi_{n_1}^2$ и $\chi_{n_2}^2$ – независимые случайные величины, распределенные по χ_n^2 с n_1 и n_2 степенями свободы.

Это распределение определяется двумя параметрами – n_1 и n_2 (рис. 1.2). Для него также разработаны таблицы.

Некоторые необходимые для проверки статистических гипотез данные об F-распределении содержатся в таблице 3 приложения 2.

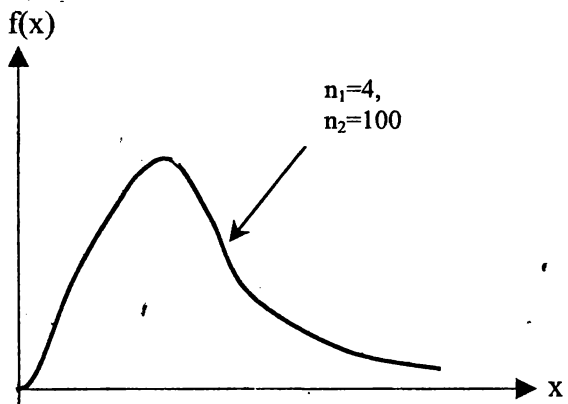


Рис. 1.2. График $f(x)$ – плотности F-распределения для $n_1 = 6$, $n_2 = 20$

1.3. Распределение выборочной дисперсии

В сумме $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n y_i^2$ слагаемые y_i имеют нормальное распределение, но зависимы, так как связаны соотношением $\sum_{i=1}^n y_i = 0$. Можно показать, что

$$\frac{nS^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} = \chi_{n-1}^2.$$

Тем самым $MS^2 = \frac{\sigma^2}{n}(n-1)$ и выборочная дисперсия S^2 является смещенной оценкой.

Смещение оценки произошло из-за того, что отклонение выборочных значений отсчитывается не от математического ожидания теоретического распределения вероятностей, которое неизвестно, а от его эмпирического аналога. Это же обстоятельство повлияло и на распределение вероятностей S^2 . Если бы мы знали значение μ математического ожидания теоретического распределения вероятностей величины ξ , из которого произведена выборка, и возводили в квадрат отклонение от него, то эмпирическая дисперсия, являясь суммой квадратов независимых нормальных величин, была бы пропорциональна с мно-

жителем $\frac{\sigma^2}{n}$ величине, имеющей распределение χ_n^2 . В силу замены а на его эмпирический аналог $\bar{X}$ с помощью замены переменных число слагаемых в сумме можно уменьшить на 1 и число степеней свободы у χ^2 распределения будет $(n-1)$.

Вообще говоря, какому бы закону ни подчинялась случайная величина, порождающая выборку, из свойств математического ожидания всегда следует

$$MS^2 = \frac{n-1}{n} D_{\xi} \neq D_{\xi}.$$

Чтобы получить несмещенную оценку, надо внести небольшую поправку: $s^2 = \frac{n}{n-1} S^2$ - оценка несмещенная;

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 - \text{для вариационного ряда общего вида};$$

$$s^2 = \frac{n}{n-1} \sum_{i=1}^k (x_i - \bar{x})^2 \frac{m_i}{n} - \text{для выборки, заданной таблицей}.$$

1.4. Интервальная оценка для дисперсии

Формула доверительного интервала для дисперсии выводится совершенно аналогично тому, как мы выводили формулу доверительного интервала для среднего, то есть используем свойства распределения вероятностей величины S^2 . Как было показано:

$$S^2 = \frac{\sigma^2}{n} \chi_{n-1}^2.$$

А для несмещенной оценки s^2 :

$$s^2 = \frac{\sigma^2}{n-1} \chi_{n-1}^2.$$

Если заданный уровень доверия равен β , то для распределения χ_{n-1}^2 по таблицам распределения χ^2 надо найти интервал, в который величина, имеющая такое распределение, попадает с вероятностью β .

В силу асимметричности распределения χ^2 надо отдельно искать правый и левый конец интервала. Очевидно, что концы интервала определяются неоднозначно. Можно строить интервал так, чтобы он имел минимальную длину. А можно строить его так, чтобы незаштрихованные области справа и слева от него имели одинаковую площадь, равную $\alpha/2$ (рис. 1.3).

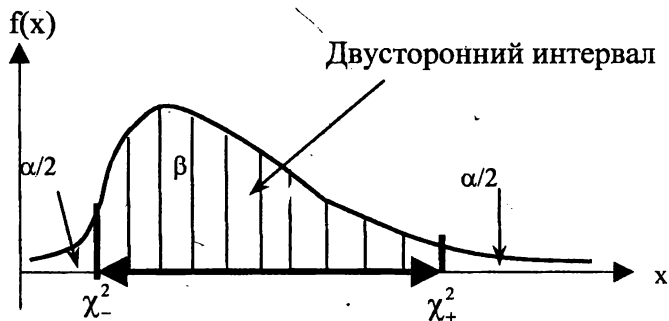


Рис. 1.3

Выпишем несколько эквивалентных уравнений:

$$P(\chi^2_- < \chi^2_{n-1} < \chi^2_+) = \beta;$$

$$P(\chi^2_- < nS^2/\sigma^2 < \chi^2_+) = \beta;$$

$$P\left(\frac{nS^2}{\chi^2_+} < \sigma^2 < \frac{nS^2}{\chi^2_-}\right) = \beta.$$

Таким образом, формула двустороннего доверительного интервала уровня доверия β для дисперсии имеет вид

$$\frac{nS^2}{\chi^2_+} < \sigma^2 < \frac{nS^2}{\chi^2_-} \quad \text{или} \quad \frac{(n-1)s^2}{\chi^2_+} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_-}.$$

Для среднеквадратического отклонения σ получаем доверительный интервал извлечением корня из границ этого интервала:

$$S\sqrt{\frac{n}{\chi^2_+}} < \sigma < S\sqrt{\frac{n}{\chi^2_-}} \quad \text{или} \quad s\sqrt{\frac{n-1}{\chi^2_+}} < \sigma < s\sqrt{\frac{n-1}{\chi^2_-}}.$$

Обычно принято пользоваться несмещенной "исправленной" оценкой для дисперсии.

В приложении 2 приведена таблица 1, в которой содержатся два конца интервала. Она позволяет получать доверительный интервал для σ^2 без каких-либо арифметических действий по заданному уровню доверия, объему выборки (из него надо вычесть единицу) и значению несмещенной оценки s^2 .

Пример 1.4. На основании выборочных наблюдений производительности труда 20 работниц было установлено, что оценка для среднего квадратического отклонения суточной выработки – 15 метров ткани. Предполагая, что производительность труда работницы имеет нормальное распределение, найти границы, в которых с надежностью 0,9 заключены генеральные дисперсия и среднеквадратическое отклонение суточной выработки работниц.

Решение задачи с помощью таблицы 1 приложения 2:

$$s^2=20/19 \cdot S^2=1,0526 \cdot 225=236,835;$$

$$236,835 \cdot 0,610 < \sigma^2 < 236,835 \cdot 1,805 \quad \text{или} \quad 144,47 < \sigma^2 < 427,487.$$

А для среднеквадратического отклонения

$$12,02 < \sigma < 20,67.$$

Еще один способ выбора интервала позволяет упростить таблицы. Выберем интервал, симметричный относительно значения S . В этом случае в таблице по вероятности β и числу степеней свободы n отыскивается значение величины $q(\beta, n)$. Доверительный интервал для σ при этом имеет вид

$$s(1-q) < \sigma < s(1+q).$$

Если $1-q < 0$, то нижняя граница заменяется на 0. То есть на самом деле эта формула имеет вид

$$\min(0, (1-q))s < \sigma < (1+q)s.$$

При отсутствии подобных таблиц можно пользоваться таблицами критических точек χ^2 распределения и приведенными выше формулами, в которых участвуют χ^2_+ и χ^2_- – критические точки для уровней значимости $\alpha/2$ и $1-\alpha/2$, где $\alpha = 1-\beta$:

$$\chi_+^2 = \chi_{\alpha/2, n-1}^2;$$

$$\chi_-^2 = \chi_{1-\alpha/2, n-1}^2.$$

Решим ту же задачу с помощью таблиц χ_n^2 распределения:

$$\alpha/2=0,05; 1-\alpha/2=0,95.$$

Согласно таблице 2 для 19-ти степеней свободы $\chi_+^2=30,1$, а $\chi_-^2=10,1$ и доверительный интервал для дисперсии:

$$20/30,1 \cdot 15^2 < \sigma^2 < 20/10,1 \cdot 15^2 \text{ или } 149,5 < \sigma^2 < 445,6$$

Для среднеквадратического отклонения

$$12,2 < \sigma < 21,1$$

Приведенный пример показывает зависимость интервала от способа построения. В то же время он демонстрирует, что интервалы очень похожи.

Таблицы χ_n^2 распределения позволяют строить односторонние интервалы для дисперсии и среднеквадратического отклонения (рис. 1.4 и 1.5).

Правосторонний интервал

$$\sigma^2 < \frac{(n-1)s^2}{\chi_{\alpha, n-1}^2}$$

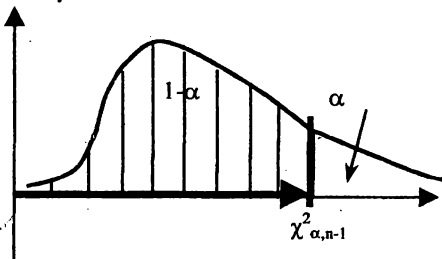


Рис. 1.4

Левосторонний интервал

$$\sigma^2 > \frac{(n-1)s^2}{\chi_{1-\alpha, n-1}^2}$$

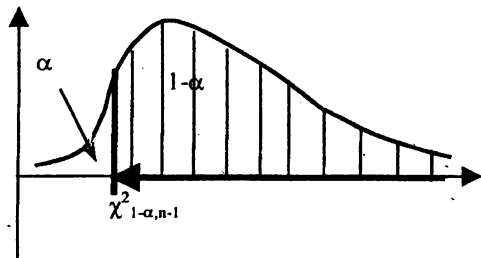


Рис. 1.5

1.5. Проверка гипотез, связанных с дисперсией

Гипотезы о дисперсиях возникают довольно часто. Например, при произвольной настройке станка может возникнуть необходимость проверки гипотезы о том, что точность изготовления деталей по некоторому параметру, задаваемая дисперсией σ^2 , равна заданной величине σ_0^2 (нулевая гипотеза H_0). Альтернативной может выступать гипотеза H_1 о том, что после ремонта станка точность изменилась, стала σ_1^2 , и новая точность просто стала другой ($\sigma_1^2 \neq \sigma_0^2$), или увеличилась ($\sigma_1^2 > \sigma_0^2$), или уменьшилась ($\sigma_1^2 < \sigma_0^2$).

1.5.1. Гипотеза о числовом значении дисперсии

Для проверки гипотез о числовом значении дисперсии при известном математическом ожидании a можно применить тот же подход, какой мы применяли для проверки числовых значений среднего. При построении критической области используется распределение вероятностей статистики, с помощью которой производится анализ.

При проверке гипотез о численном значении среднего при известной дисперсии использовалась статистика $T = \frac{\bar{x} - a_0}{\frac{\sigma}{\sqrt{n}}}$, имеющая стандартное нормальное распределение $N(0,1)$.

Для поиска границ критических областей использовались таблицы нормального распределения.

При проверке гипотез о численном значении среднего при неизвестной дисперсии использовалась статистика $T = \frac{\bar{x} - a_0}{\frac{S}{\sqrt{n-1}}}$, имеющая

распределение Стьюдента с $(n-1)$ степенью свободы. Для поиска границ критических областей использовались таблицы распределения Стьюдента.

При проверке гипотез о численном значении дисперсии ($\sigma = \sigma_0$) при неизвестном среднем a используется статистика: $\chi^2 = \frac{nS^2}{\sigma_0^2}$, имеющая распределение χ^2_{n-1} . Следовательно, для поиска границ критических областей используются таблицы распределения Пирсона (χ^2).

Критерии проверки для гипотезы H_0 , состоящей в том, что $\sigma = \sigma_0$, по статистике χ^2 для уровня значимости α приведены в табл. 1.6. (рис. 1.6-1.8).

Таблица 1.6

Альтернативная гипотеза	Критерий отклонения гипотезы H_0 по статистике $\chi^2 = \frac{nS^2}{\sigma_0^2}$
$\sigma^2 = \sigma_1^2 > \sigma_0^2$	$\chi^2 > \chi^2_{\alpha, n-1}$
$\sigma^2 = \sigma_1^2 < \sigma_0^2$	$\chi^2 < \chi^2_{1-\alpha, n-1}$
$\sigma^2 = \sigma_1^2 \neq \sigma_0^2$	$\chi^2 > \chi^2_{\alpha/2, n-1}$ либо $\chi^2 < \chi^2_{1-\alpha/2, n-1}$

Граница критической области $\chi^2_{\alpha, k}$ берется из таблицы критических точек распределения χ^2 (табл. 2 приложения 2).

Односторонние проверки

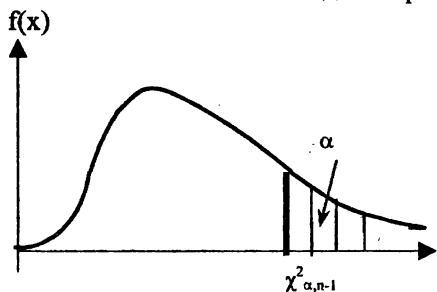


Рис. 1.6

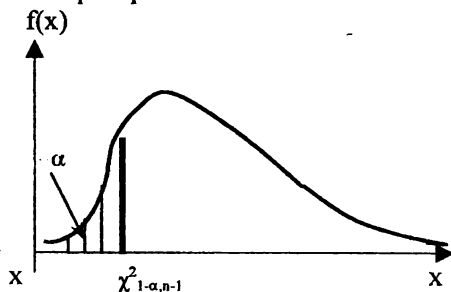


Рис. 1.7

Двусторонняя проверка

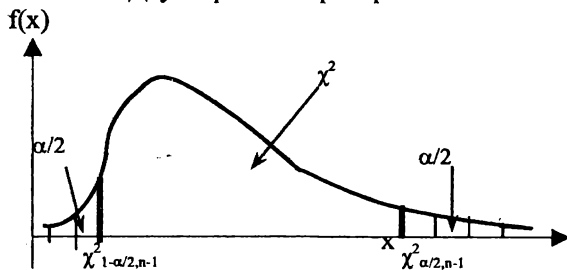


Рис. 1.8

Пример 1.5. По данным примера 1.4 о суточной выработке работниц на уровне значимости $\alpha=0,1$ проверить гипотезу о том, что среднеквадратическое отклонение суточной выработки работниц равно 20 метрам.

Решение. Проверяемая гипотеза $H_0: \sigma^2=20^2=400$. Конкурирующая гипотеза $H_1: \sigma^2 \neq 400$. Откуда следует, что доверительный интервал для σ^2 : $149,5 < \sigma^2 < 445,6$. Так как значение $\sigma^2=20^2=400$ принадлежит этому интервалу, то на уровне значимости $\alpha=0,1$ гипотеза H_0 не отвергается, то есть имеющиеся данные не противоречат предположению о том, что среднее квадратическое отклонение суточной выработки работниц равно 20 метрам.

Покажем, как решается аналогичная задача без привлечения понятия доверительного интервала.

Пример 1.6. Точность работы станка-автомата проверяется по дисперсии контролируемого размера изделий, которая не должна превышать 0,15. По пробе из 25 случайно отобранных изделий вычислена оценка дисперсии: $S^2=0,25$. При уровне значимости $\alpha=0,05$ дать ответ на вопрос: обеспечивает ли станок требуемую точность? Предполагается, что размер изделия – нормально распределенная случайная величина.

Решение. Станок будет обеспечивать требуемую точность, если генеральная дисперсия $\sigma^2 < 0,15$. Поскольку выборочная дисперсия $S^2=0,25 > 0,15$, примем за нулевую гипотезу $H_0: \sigma^2=0,15$, а за альтернативную $H_1: \sigma^2 > 0,15$. В таблице 2 отыскиваем $\chi^2_{0,05;24}=36,4$.

Таким образом, критическая область $(36,4; +\infty)$.

Вычисляем значение статистики $\chi^2 = \frac{nS^2}{\sigma_0^2} = \frac{25 \cdot 0,25}{0,15} = 41,7$. Так

как значение статистики попало в критическую область, гипотезу H_0 отвергаем, то есть считаем, что станок не обеспечивает требуемую точность.

1.5.2. Сравнение дисперсий двух совокупностей

Пусть имеются две нормально распределенные совокупности, дисперсии которых равны σ_1^2 и σ_2^2 . Необходимо проверить нулевую гипотезу о равенстве дисперсий, то есть $H_0: \sigma_1^2 = \sigma_2^2$ относительно конкурирующей $H_1: \sigma_1^2 \neq \sigma_2^2$, или $\sigma_1^2 > \sigma_2^2$, или $\sigma_1^2 < \sigma_2^2$. Для проверки H_0 из этих совокупностей взяты две независимые выборки объемом n_1 и n_2 . Для оценки дисперсий σ_1^2 и σ_2^2 используются “исправленные” выборочные дисперсии s_1^2 и s_2^2 .

Вспомним, что

$$s^2 = \frac{\sigma^2}{n-1} \chi_{n-1}^2.$$

Отсюда следует, что при справедливости гипотезы $\sigma_1^2 = \sigma_2^2 = \sigma^2$ случайная величина F , определяемая соотношением

$$F = \frac{s_1^2}{s_2^2},$$

имеет F -распределение Фишера-Снедекора с $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$ степенями свободы. Следовательно, для поиска границ критических областей используются таблицы F -распределения Фишера-Снедекора.

Критерии проверки для гипотезы H_0 , состоящей в том, что $\sigma_1^2 = \sigma_2^2$, по статистике F для уровня значимости α приведены в табл. 1.7.

Таблица 1.7

Альтернативная гипотеза	Критерий отклонения гипотезы H_0 по статистике $F = \frac{s_1^2}{s_2^2}$
$\sigma^2 = \sigma_1^2 > \sigma_2^2$	$F > F_{\alpha, n_1-1, n_2-1}$
$\sigma^2 = \sigma_1^2 \neq \sigma_2^2$	$F > F_{\alpha/2, n_1-1, n_2-1}$ либо $F < F_{1-\alpha/2, n_1-1, n_2-1}$ ($F < \frac{1}{F_{\alpha/2, n_2-1, n_1-1}}$)

Граница критической области F_{α, n_1-1, n_2-1} берется из таблицы критических точек распределения F (таблица 3). Для отыскания значений $F_{1-\alpha/2, n_1-1, n_2-1}$ надо воспользоваться равенством (очевидным, если выборки поменять местами):

$$\text{для } P > 0,5 \quad F_P(k_1, k_2) = \frac{1}{F_{1-P}(k_2, k_1)}.$$

Пример 1.7. На двух токарных станках обрабатываются втулки. Отобраны две пробы: из втулок, сделанных на первом станке $n_1 = 15$ штук и на втором станке $n_2 = 18$ штук. По данным этих выборок рассчита-

ны выборочные дисперсии: $S_1^2=8,5$ (для первого станка) и $S_2^2=6,3$ (для второго станка). Полагая, что размеры втулок подчиняются нормальному закону распределения, на уровне значимости $\alpha=0,05$ выяснить, можно ли считать, что станки обладают различной точностью (точность станка – это средний разброс, то есть требуется проверить гипотезу о равенстве дисперсий).

Решение. $H_0: \sigma_1^2=\sigma_2^2$, то есть дисперсии размера втулок для обоих станков одинаковы. В качестве конкурирующей гипотезы возьмем $H_1: \sigma_1^2>\sigma_2^2$ (дисперсия больше для первого станка). Статистика критерия:

$$F = \frac{\frac{n_1}{n_1-1} S_1^2}{\frac{n_2}{n_2-1} S_2^2} = \frac{(15/14) \cdot 8,5}{(18/17) \cdot 6,3} = 1,37.$$

По таблице 3 приложения 2 критическое значение F-критерия на уровне значимости $\alpha=0,05$ при числе степеней свободы $k_1=n_1-1=14$ и $k_2=n_2-1=17$, то есть $F_{0,05;14;17}=2,33$. Так как $F < F_{0,05;14;17}$ ($1,37 < 2,33$), то гипотеза H_0 не отвергается, то есть имеющиеся данные не позволяют считать, что станки обладают различной точностью.

Если в качестве конкурирующей гипотезы в данной задаче взять $H_1: \sigma_1^2 \neq \sigma_2^2$, то гипотеза H_0 отвергается, если

$$F > F_{0,025;14;17} \text{ либо } F < F_{0,975;14;17} \left(F < \frac{1}{F_{0,025;17;14}} \right).$$

При проверке гипотезы о равенстве средних мы предполагали, что генеральные дисперсии для обеих выборок совпадают. Так что корректно сначала проверить с помощью F-критерия гипотезу о совпадении генеральных дисперсий и только после того, как эта гипотеза подтвердится, выполнить проверку гипотезы о совпадении генеральных средних.

В главе 3 мы рассмотрим критерий Бартлета, который используется для проверки гипотезы о совпадении нескольких дисперсий.

В заключение отметим следующее. Выше при проверке гипотез предполагалась нормальность распределения исследуемых случайных величин. Однако специальные исследования показали, что предложенные алгоритмы весьма устойчивы (особенно при больших объемах выборок) по отношению к отклонению от нормального распределения.

1.6. Оценка необходимого числа опытов для сравнения двух дисперсий

Пусть мы хотим проверить гипотезу о равенстве двух дисперсий при сравнении с альтернативой $\sigma_1^2 > \sigma_2^2$ при уровне значимости α и хотим, чтобы ошибка 2-го рода при этом не превосходила некоторого значения β . Если мы хотим, чтобы вероятности обеих ошибок – и 1-го, и 2-го рода – не превосходили α и β соответственно, число испытаний должно быть достаточно большим. Не приводя доказательств, дадим формулу для минимального числа испытаний для самого естественно-го случая: $n_1 = n_2$ и $\alpha = \beta$.

Обозначим через Θ альтернативное единице значение отношения:

$$\frac{\sigma_1^2}{\sigma_2^2} = \Theta$$

(для гипотезы H_0 это отношение равно единице, для гипотезы H_1 это отношение равно Θ).

Минимальное число опытов, при котором будут соблюдены требования к величине ошибок 1-го и 2-го рода, определяется с помощью формулы:

$$F_{\alpha}(f, f) / F_{1-\alpha}(f, f) = \Theta,$$

где $f = n - 1$ – число степеней свободы.

Принимая во внимание, что $F_{1-\alpha}(f, f) = 1 / F_{\alpha}(f, f)$, надо найти такое n , при котором выполняется: $F_{\alpha}^2(n, n) = \Theta$. Такое n отыскивается на диагонали таблицы F-распределения. Например, пусть $\alpha = \beta = 0,05$, $\Theta = 4$ (проверяем гипотезу о том, что одно среднеквадратическое отклонение вдвое больше, чем второе на уровне значимости 0,05). В таблице 3 отыскиваем n , для которого $F_{0,05}(n, n) = 2$. Таким n будет $n = 24$. Следовательно, нужно каждым методом произвести $n = 25$ измерений.

2. ПРОВЕРКА ГИПОТЕЗ О ЗАКОНЕ РАСПРЕДЕЛЕНИЯ

До сих пор в наших рассуждениях вид закона распределения генеральной совокупности предполагался известным. Неизвестными считались только параметры распределения, для которых мы и находили статистические оценки. Действительно, предположение о виде закона распределения может быть выдвинуто исходя из теоретических предпосылок. Например, выполнение условий центральной предельной теоремы (очень грубо – случайная величина является суммой большого числа независимых случайных величин) может свидетельствовать о нор-

мальном законе распределения случайной величины. Наконец, гистограмма дает основания сделать вывод о виде распределения. Параметры распределения, как правило, неизвестны, поэтому их заменяют эмпирическими оценками по выборке. Но между эмпирическим и теоретическим распределениями неизбежны расхождения. Естественно возникает вопрос: объясняются ли эти расхождения только случайными обстоятельствами, связанными с ограниченным числом наблюдений, или они являются существенными и связаны с тем, что теоретический закон распределения подобран неудачно? Ответ на этот вопрос дают **критерии согласия** (согласие гипотезы о теоретическом распределении с опытными данными).

Пусть необходимо проверить нулевую гипотезу H_0 о том, что исследуемая случайная величина имеет определенный закон распределения. Для проверки гипотезы H_0 выбирают некоторую случайную величину U , характеризующую степень расхождения теоретического и эмпирического закона распределения. Зная закон распределения величины U , можно найти вероятность того, что U приняла значение не меньшее, чем фактически наблюдаемое в опыте u . Если вероятность $P(U > u) = \alpha$ мала, то в соответствии с принципом практической уверенности это означает, что такие как в опыте и большие отклонения практически невозможны. В этом случае гипотезу H_0 отвергают. Если же вероятность $P(U > u) = \alpha$ не мала, то расхождение между эмпирическим и теоретическим распределениями несущественно и гипотезу H_0 можно считать правдоподобной или по крайней мере не противоречащей опытными данным.

2.1. Критерий согласия Пирсона (χ^2 -критерий)

Чаще всего в качестве меры расхождения U берется величина χ^2 , которая вычисляется следующим образом:

$$U = \chi^2 = \sum_{i=1}^m \frac{(n_i - np_i)^2}{np_i},$$

где n – объем выборки.

Если генеральное распределение дискретно, то p_i – это вероятности принять значение x_i , а n_i – эмпирические частоты для значения x_i (сколько раз значение x_i встретилось в выборке). Если генеральное распределение непрерывно или различных принимаемых значений очень много, то выборка группируется. Для интервалов группировки выборки (x_i, x_{i+1}) подсчитываются частоты n_i (число попаданий данных опыта в интервал) и теоретические вероятности p_i попаданий в интервал (x_i, x_{i+1}) .

Для вычисления вероятностей p_i надо воспользоваться таблицами генерального распределения: $p_i = F(x_{i+1}) - F(x_i)$.

Если мы поделим числитель и знаменатель каждого слагаемого на n^2 , то получим сумму квадратов величин, каждая из которых согласно теореме Муавра-Лапласа распределена приблизительно нормально:

$$U = \chi^2 = \sum_{i=1}^m \frac{(n_i / n - p_i)^2}{p_i / n}.$$

Величины не являются независимыми, так как имеет место связь $\sum p_i = 1$. Кроме этой связи, имеется связь и между p_i в случае, если параметры теоретического распределения неизвестны и их заменяют эмпирическими оценками. Это уменьшает число степеней свободы у распределения величины U .

Доказано, что статистика U при $n \rightarrow \infty$ имеет χ^2 распределение с $k = m - r - 1$ степенями свободы, где m – число слагаемых в сумме (интервалов группировки вариационного ряда); r – число параметров теоретического распределения, вычисленных по экспериментальным данным.

Необходимо, чтобы в каждом интервале было по крайней мере пять наблюдений. Если для каких-то интервалов это правило не соблюдается, необходимо объединить соседние интервалы.

Числа n_i и $n'_i = np_i$ называются соответственно эмпирическими и теоретическими частотами.

Схема применения χ^2 -критерия следующая:

1. Вычисляется значение статистики U .
2. Для выбранного уровня значимости α по таблице χ^2 -распределения находят критическое значение $\chi^2_{\alpha, k}$ для числа степеней свободы $k = m - r - 1$.
3. Если $U > \chi^2_{\alpha, k}$, то гипотеза H_0 отвергается; если $U < \chi^2_{\alpha, k}$, то гипотеза H_0 не противоречит опытным данным.

Пример 2.1 (параметры теоретического распределения известны).

При 4000 бросаниях монеты Бюффон получил $m=2048$ выпадений герба. Согласуется ли это с гипотезой о том, что существует постоянная вероятность $p=0,5$ выпадения герба?

Решение. Примем, что выпадение решки соответствует значению $x_1=0$, а выпадение герба – значению $x_2=1$. Занесем теоретические и эмпирические данные в расчетную таблицу (табл. 2.1).

$$U = \chi^2 = 1,152 + 1,152 = 2,304.$$

Число степеней свободы k равно 1, так как принимаются всего два

Таблица 2.1

x_i	n_i	n'_i	$n_i - n'_i$	$(n_i - n'_i)^2$	$(n_i - n'_i)^2 / n'_i$
0	1952	2000	-48	2304	1,152
1	2048	2000	48	2304	1,152

значения, следовательно, $m=2$, неизвестных параметров у распределения нет, так как $p=0,5$ и $g=0$. Тем самым $k = m - r - 1 = 1$. По табл. 2 χ^2 -распределения находим, что гипотеза о существовании постоянной вероятности $p=0,5$ выпадения герба хорошо согласуется с опытными данными (например, на уровне значимости $0,05$ $\chi^2_1=3,84$ и $u<3,84$).

Пример 2.2 (параметры теоретического распределения неизвестны).

Отдел технического контроля проверил $n=200$ партий одинаковых изделий и получил следующее эмпирическое распределение (в первой строке указано количество x_i нестандартных изделий в одной партии; во второй строке – частота n_i , то есть количество партий, содержащих x_i нестандартных изделий):

x_i	0	1	2	3	4
n_i	116	56	22	4	2

Требуется при уровне значимости $0,05$ проверить гипотезу о том, что число нестандартных изделий X распределено по закону Пуассона.

Решение.

1. Найдем выборочную среднюю $\bar{X}$:

$$\bar{X} = (116 \cdot 0 + 56 \cdot 1 + 22 \cdot 2 + 4 \cdot 3 + 2 \cdot 4) / 200 = 0,6.$$

2. Примем в качестве оценки λ распределения Пуассона выборочную среднюю $\lambda=0,6$ и, положив $i = 0, 1, 2, 3, 4$, найдем по табл. 4 вероятности p_i появления i нестандартных изделий в 200 партиях: $p_0 = 0,5488$; $p_1 = 0,3293$; $p_2 = 0,0988$; $p_3 = 0,0198$; $p_4 = 0,0030$.

3. Найдем теоретические частоты по формуле

$$n'_i = np_i = 200p_i$$

и подставим эти значения в расчетную таблицу (табл. 2.2); в расчетной таблице объединяем значения малочисленных частот $4+2=6$ и соответствующих им теоретических частот $3,96 + 0,60 = 4,56$.

Таблица 2.2

i	n_i	n'_i	$n_i - n'_i$	$(n_i - n'_i)^2$	$(n_i - n'_i)^2 / n'_i$
0	116	109,76	6,24	38,9376	0,3548
1	56	65,86	-9,86	97,2196	1,4762
2	22	19,76	2,24	5,0176	0,2539
3	6	4,56	1,44	2,0736	0,4547
Σ	200				$\chi^2=2,54$

4. Вычисляем число степеней свободы $k = 4 - 1 - 1 = 2$ (четыре значения, один параметр λ заменен эмпирическим значением) и отыскиваем в табл. 2 соответствующее уровню значимости 0,05 критическое значение. Оно равно 6.

Так как $\chi^2_{\text{набл}} < \chi^2_{\text{крит}} (2,54 < 6)$, то нет оснований отвергнуть гипотезу о распределении случайной величины X по закону Пуассона.

2.2. Проверка гипотезы о том, что генеральное распределение нормальное

1. *Параметры нормального распределения μ и σ известны.* В этом случае удобно взять в качестве концов интервалов группировки точки: $\mu - 3\sigma$, $\mu - 2\sigma$, $\mu - \sigma$, μ , $\mu + \sigma$, $\mu + 2\sigma$, $\mu + 3\sigma$. Теоретические вероятности попадания в эти интервалы в соответствии со свойствами нормального распределения соответственно равны: 0,0217; 0,137; 0,34; 0,34; 0,137; 0,0217 (так как $P\{\mu - 3\sigma < X < \mu + 3\sigma\} = 0,9973$; $P\{\mu - 2\sigma < X < \mu + 2\sigma\} = 0,954$; $P\{\mu - \sigma < X < \mu + \sigma\} = 0,68$). При $n < 250$ теоретические частоты попадания в крайние интервалы слишком малы и их надо объединить с соседними.

2. *Параметры нормального распределения μ и σ неизвестны.* В этом случае вычисляются для μ и σ оценки $\bar{x}$ и S^2 . Для расчета теоретических вероятностей p_i попадания случайной величины в интервал $[x_i, x_{i+1}]$ воспользуемся тем, что

$$p_i(x_i \leq X \leq x_{i+1}) = \frac{1}{2} \left[\Phi\left(\frac{x_{i+1} - \mu}{\sigma}\right) - \Phi\left(\frac{x_i - \mu}{\sigma}\right) \right] = F\left(\frac{x_{i+1} - \mu}{\sigma}\right) - F\left(\frac{x_i - \mu}{\sigma}\right),$$

где $\Phi(x)$ – функция Лапласа, а $F(x)$ – функция распределения стандартной нормальной величины. Следовательно, можно использовать таблицы этих функций. Шапка расчетной таблицы будет иметь вид:

Интервал [x_i, x_{i+1}]	Эмпирические частоты n_i	Вероят- ности p_i	Теоретические частоты $n'_i = np_i$	$(n_i - n'_i)^2$	$(n_i - n'_i)^2 / n'_i$
--------------------------------	-------------------------------	------------------------	--	------------------	-------------------------

Пример 2.3. Вспомним таблицу выработки рабочих из юниты 3 (табл. 2.3).

Таблица 2.3

Выработка рабочих в процентах к предыдущему году

Середина i-го ин- тервала	Выработка в процентах	Количество рабочих (частота)	Доля рабо- чих	Накоплен- ная частота	Накопленная относительная частота
97	94-100	3	0,03	3	0,03
103	100-106	7	0,07	3+7=10	0,10
109	106-112	11	0,11	10+11=21	0,21
115	112-118	20	0,20	21+20=41	0,41
121	118-124	28	0,28	41+28=69	0,69
127	124-130	19	0,19	69+19=88	0,88
133	130-136	10	0,10	88+10=98	0,98
139	136-142	2	0,02	98+2=100	1,00

Представленная в табл. 2.3 **накопленная частота** показывает, сколько наблюдалось вариантов со значением признака, меньшим x . Отношение накопленной частоты к общему числу наблюдений мы называли **накопленной относительной (эмпирической) частотой**. К этим понятиям мы еще вернемся. На примере этой таблицы покажем, как ведется вычисление значения статистики χ^2 .

Взяв в качестве наблюдаемых значений середины интервалов, вычисляем $\bar{x}$ и S^2 :

$$\bar{x} = 119,2;$$

$$S^2 = 87,48.$$

На примере первого интервала [94–100] покажем, как вычисляются p_i :

$$\begin{aligned} p_i(94 < X < 100) &= 1/2[\Phi((100 - 119,2) / 9,35) - \Phi((94 - 119,2) / 9,35)] = \\ &= 1/2[\Phi(-2,05) - \Phi(-2,69)] = 1/2(-0,9596 + 0,9928) = 0,0166. \end{aligned}$$

Учитывая, что в рассматриваемом эмпирическом распределении частоты первого и последнего интервалов ($n_1=3$, $n_8=2$) меньше 5, надо объединить эти интервалы с соседними (табл. 2.4).

Таблица 2.4

Интервал [x_i , x_{i+1}]	Эмпириче- ские частоты n_i	Вероятности p_i	Теоретические частоты $n'_i=np_i$	$(n_i - n'_i)^2$	$(n_i - n'_i)^2 / n'_i$
94-106	3+7=10	0,017+0,059 =0,076	1,7+5,9=7,1	5,76	0,758
106-112	11	0,141	14,1	9,61	0,682
112-118	20	0,228	22,8	7,84	0,344
118-124	28	0,247	24,7	10,89	0,441
124-130	19	0,182	18,2	0,64	0,035
130-142	10+2=12	0,087+0,029 =0,116	8,7+2,9=11,6	0,16	0,014
Σ	100				2,27

Число степеней свободы $k = 6 - 2 - 1 = 3$.

Теперь мы можем решить обычную задачу: на уровне значимости 0,05 проверить гипотезу о том, что выработка рабочих распределена нормально. Критическое значение для уровня значимости 0,05 равно 7,82 (см. табл. 2 приложения 2). Так как $\chi^2_{\text{набл}} < \chi^2_{\text{крит}}$ ($2,27 < 7,82$), то нет оснований отвергнуть гипотезу о нормальном распределении случайной величины X .

Сложных вычислений при проверке гипотезы о нормальном распределении можно избежать. Для этого воспользуемся значениями накопленных частот. Составим расчетную таблицу по схеме:

Значение x_i	Частота n_i	Накопленная частота $N_i = \sum_{r=1}^i n_r - 0,5$	Относительная накопленная частота N_i/n	Квантили u_{pi}
-------------------	------------------	---	---	----------------------

Значения квантилей находим в табл. 6 по значениям N_i/n .

Затем в прямоугольной системе координат строим точки $(x_1, u_1), (x_2, u_2), \dots, (x_k, u_k)$. Если эти точки лежат вблизи некоторой прямой, то нет оснований отвергнуть гипотезу о нормальном распределении. Если гипотеза о нормальном законе справедлива, уравнение этой пря-

мой $u = (x - \bar{x})/S$, то есть в качестве оценки математического ожидания a можно принять абсциссу точки пересечения построенной прямой с осью Ox . В качестве оценки среднего квадратического отклонения σ можно принять разность абсцисс этой точки и точки пересечения построенной линии с прямой $u=-1$ (рис. 2.1).

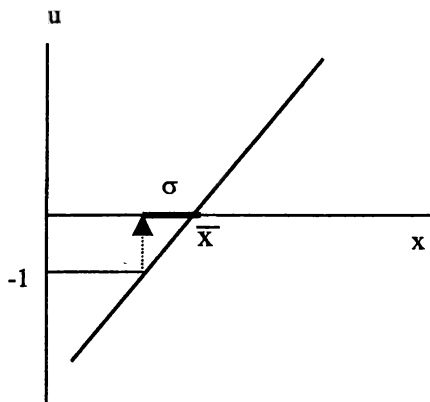


Рис. 2.1

Желающие могут в качестве упражнения провести соответствующие расчеты и построить эту прямую для выработки продавцов.

2.3. Проверка гипотезы о равномерном распределении

1. Концы интервала a и b известны. Теоретические вероятности попадания в интервал $[x_i, x_{i+1}]$ равны: $p_i = (x_{i+1} - x_i)/(b-a)$. В естественном случае, когда все интервалы группировки имеют одинаковую длину, $p_i = 1/k$, где k – число интервалов группировки.

Число степеней свободы равно числу интервалов группировки минус 1.

2. Концы интервала a и b неизвестны. Сначала вычисляются оценки для a и b . Эти оценки находим с помощью метода моментов:

$$M(X) = (a + b) / 2;$$

$$\sigma(X) = (b - a) / 2\sqrt{3}.$$

Следовательно, чтобы найти оценки для a и b , надо решить систему:

$$(a' + b') / 2 = \bar{x};$$

$$(b' - a') / 2\sqrt{3} = S.$$

Решив эту систему, получим:

$$a' = \bar{x} - \sqrt{3}S, \quad b' = \bar{x} + \sqrt{3}S.$$

Далее найдем плотность распределения: $f(x) = \frac{1}{b' - a'}.$

Теоретические частоты ищем по формулам:

$$n'_1 = n \cdot f(x) \cdot (x_1 - a');$$

$$n'_2 = n \cdot f(x) \cdot (x_2 - x_1);$$

.....

$$n'_k = n \cdot f(x) \cdot (b' - x_k).$$

Далее с помощью стандартной расчетной таблицы находим значение критерия χ^2 . Число степеней свободы равно числу интервалов группировки минус 3.

Замечание. Согласно методу минимакса (наименьшего максимума абсолютных отклонений (см., например, работу [5]) в качестве оценки для a и b можно взять $\min x_i$ и $\max x_i$ соответственно. Это значительно упрощает все расчеты. Интервалы группировки, естественно, можно взять одинаковой длины: $x_1 = \min x_i$; $x_{k+1} = \max x_i$; $p_i = 1/k$, где k – число интервалов группировки.

2.4. Критерий Колмогорова

На практике для проверки гипотезы о виде распределения вероятностей кроме критерия χ^2 часто применяют **критерий Колмогорова**, в котором при расчете меры расхождения между теоретическим и эмпирическим распределениями используют максимальное значение абсолютной величины разности между эмпирической функцией распределения $F_n(x)$ и соответствующей теоретической функцией распределения $F(x)$:

$$D = \max |F_n(x) - F(x)|.$$

Мы уже упоминали, что кумулята (или эмпирическая функция распределения) в каждой точке сходится по вероятности к теоретической функции распределения.

Теорема, доказанная Колмогоровым, позволяет уточнить смысл последнего высказывания и задает критерий для проверки гипотезы о теоретическом распределении вероятностей. Согласно теореме Колмогорова D (наибольшая величина отклонения эмпирической функции от теоретической) имеет порядок $\frac{c}{\sqrt{n}}$. Точнее, D – случайная величина. В теореме доказывается, что при неограниченном увеличении числа наблюдений ($n \rightarrow \infty$)

$$P\left\{\max_x |F_n(x) - F(x)| < \frac{t}{\sqrt{n}}\right\} \xrightarrow{n \rightarrow \infty} K(t),$$

где $K(t)$ – так называемое распределение Колмогорова. Взяв, например, $t=1,36$, из таблиц функции Колмогорова (таблица 10 приложения 2) получаем, что при достаточно большом объеме выборки неравенство

$|F_n(x) - F(x)| < \frac{1,36}{\sqrt{n}}$ (или $\sqrt{n} |F_n(x) - F(x)| < 1,36$) выполняется с

вероятностью 0,95 (т.е. в среднем для 95% выборок). Таким образом,

если взять величину $\lambda = \sqrt{n} |F_n(x) - F(x)|$ в качестве меры расхождения между теоретическим и эмпирическим распределением (она называется **статистикой критерия Колмогорова**), можно использовать то, что при больших n она имеет распределение Колмогорова.

Критерий Колмогорова достаточно часто применяется на практике благодаря своей простоте. Однако в принципе его применение возможно лишь тогда, когда теоретическая функция распределения $F(x)$ задана полностью. Но такой случай на практике встречается весьма редко. Обычно из теоретических соображений известен лишь вид функции распределения, а ее параметры определяются по эмпирическим данным. При применении критерия χ^2 это обстоятельство учитывается соответствующим уменьшением числа степеней свободы. Такого рода поправок в критерии Колмогорова не предусмотрено. В результате есть риск принять нулевую гипотезу H_0 о законе распределения как правдоподобную, в то время как на самом деле она противоречит опытным данным.

В качестве примера рассмотрим те же данные о выработке рабочих, хотя в данном случае параметры распределения неизвестны:

x	94	100	106	112	118	124	130	136	142
$F_n(x)$	0,010	0,030	0,100	0,210	0,410	0,690	0,880	0,980	1,00
$F(x)$	0,004	0,021	0,080	0,221	0,449	0,695	0,878	0,964	0,993

$D = |F_n(118) - F(118)| = |0,410 - 0,449| = 0,039$, следовательно, $\sqrt{n} \cdot D = 0,39$. Для уровня значимости 0,05 критическое значение критерия Колмогорова равно 1,36. Так как наблюдаемое значение 0,39 оказалось меньше критического значения 1,36, можно считать, что гипотеза H_0 не противоречит опытным данным (значение 1,36 наряду с правилами 2σ и 3σ следует помнить наизусть).

2.5. Проверка гипотез об однородности выборок

Гипотезы об однородности выборок – это гипотезы о том, что рассматриваемые выборки извлечены из одной и той же генеральной совокупности.

Пусть имеются две независимые выборки, произведенные из генеральных совокупностей с неизвестными теоретическими функциями распределения $F_1(x)$ и $F_2(x)$. Проверяемая нулевая гипотеза имеет вид $H_0: F_1(x) = F_2(x)$ против конкурирующей $H_1: F_1(x) \neq F_2(x)$. Будем предполагать, что функции $F_1(x)$ и $F_2(x)$ непрерывны.

Критерий Колмогорова-Смирнова использует ту же самую идею, что и критерий Колмогорова, но только по критерию Колмогорова-Смирнова сравниваются две эмпирические функции распределения.

Статистика критерия Колмогорова-Смирнова имеет вид

$$\lambda' = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot \max |F_{n_1}(x) - F_{n_2}(x)|,$$

где $F_{n_1}(x)$ и $F_{n_2}(x)$ – эмпирические функции распределения, построенные по двум выборкам объемов n_1 и n_2 .

Гипотеза H_0 отвергается, если фактически наблюдаемое значение статистики λ' больше $\lambda'_{кр}$, и принимается в противном случае. При $n_1 > 50$ и $n_2 > 50$ статистика λ' имеет практически распределение Колмогорова. При меньших n надо пользоваться специальными таблицами.

Пример 2.4. В течение месяца выборочно осуществлялась проверка торговых точек города по продаже овощей. Результаты двух проверок по недовесам покупателям одного вида овощей приведены в табл. 2.5.

Можно ли считать, что на уровне значимости 0,05 недовесы овощей описываются одной и той же функцией распределения?

Решение. Вычислим накопленные частоты и, поделив их на n_1 и n_2 соответственно, получим значения эмпирических функций распределения для обеих проверок. Данные запишем в табл. 2.6.

Таким образом, $\lambda' = \sqrt{\frac{110 \cdot 100}{110 + 100}} \cdot 0,089 = 0,644$.

Так как $\lambda' < \lambda_{кр}$ ($0,644 < 1,36$), нулевая гипотеза не отвергается, т.е. недовесы покупателям являются устойчивым и закономерным процессом при продаже овощей в данном городе.

Таблица 2.5

Интервалы недовесов, г	Частоты для выборки 1	Частоты для выборки 2
0-10	3	5
10-20	10	12
20-30	15	8
30-40	20	25
40-50	12	10
50-60	5	8
60-70	25	20
70-80	15	7
80-90	5	5
	$n_1=110$	$n_2=100$

Таблица 2.6

x_i	$\Pi_{i,1}$ накопл.	$\Pi_{i,2}$ накопл.	$F_{n1}(x_i)$	$F_{n2}(x_i)$	$ F_{n1}(x_i) - F_{n2}(x_i) $
10	3	5	$3/110=0,027$	$5/100=0,050$	0,023
20	$10+3=13$	17	$13/110=0,118$	0,170	0,052
30	$13+15=28$	25	0,254	0,250	0,004
40	$28+20=48$	50	0,436	0,500	0,064
50	$48+12=60$	60	0,545	0,600	0,055
60	$60+5=65$	68	0,591	0,680	0,089
70	$65+25=90$	88	0,818	0,880	0,072
80	$90+15=105$	95	0,955	0,950	0,005
90	$105+5=110$	100	1,000	1,000	0,000

2.6. Проверка гипотез об однородности выборок с помощью критерия χ^2

Применение критерия χ^2 позволяет проверить гипотезу о совпадении распределений вероятностей для любого числа выборок.

Пусть имеется k независимых выборок объемом n_j каждая ($j=1,2,\dots,k$). Данные каждой выборки сгруппированы в m одинаковых интервалов. Частоту попадания объектов j -й выборки в i -й интервал обозначим через n_{ij} (табл. 2.7).

Таблица 2.7

$i \setminus j$	1	2	...	k	$\sum_{j=1}^k n_{ij}$	p_i
1	n_{11}	n_{12}	...	n_{1k}	n_{1*}	n_{1*}/n
2	n_{21}	n_{22}	...	n_{2k}	n_{2*}	n_{2*}/n
...	...	...	...	...	...	...
m	n_{m1}	n_{m2}	...	n_{mk}	n_{m*}	n_{m*}/n
$\sum_{i=1}^m n_{ij}$	n_{*1}	n_{*2}	...	n_{*k}	n	1

Объединим все выборки в одну и вычислим для нее частоты $n_{i\text{общ}} = n_{i*} = n_{i1} + n_{i2} + \dots + n_{ik}$ попадания в i -й интервал и число испытаний для этой выборки $n = n_{*1} + n_{*2} + \dots + n_{*k}$. Примем в качестве теоретической частоты попадания в i -й интервал частоту, получаемую при объединении всех выборок в одну, то есть величину

$$p_i = (n_{i1} + n_{i2} + \dots + n_{ik}) / (n_{*1} + n_{*2} + \dots + n_{*k}) = n_{i\text{общ}} / n.$$

Затем для каждой выборки вычислим значение критерия χ^2 , как если бы мы проверяли гипотезу, что распределение этой выборки совпадает с сосчитанным теоретическим распределением, и сложим между собой значения критериев для всех выборок. Сумма этих случайных величин, естественно, будет иметь распределение χ^2 . Число степеней свободы равно $(m-1)(k-1)$.

Таким образом, для проверки гипотезы об однородности k выборок надо вычислить значение статистики:

$$\chi_n^2 = \sum_{j=1}^k \sum_{i=1}^m \frac{(n_{ij} - n_j \cdot p_i)^2}{n_j \cdot p_i} = n \sum_{j=1}^k \sum_{i=1}^m \frac{\left(n_{ij} - 1/n \cdot \left(\sum_{j=1}^k n_{ij} \cdot \sum_{i=1}^m n_{ij} \right) \right)^2}{\sum_{j=1}^k n_{ij} \cdot \sum_{i=1}^m n_{ij}},$$

где $p_i = \frac{1}{n} \sum_{j=1}^k n_{ij}$; $n = \sum_{j=1}^k \sum_{i=1}^m n_{ij}$,

и воспользоваться таблицей критических точек распределения χ^2 для числа степеней свободы $(m - 1)(k - 1)$.

Пример 2.5. В табл. 2.8 приведено распределение служащих двух отраслей по зарплате. Необходимо проверить гипотезу однородности распределений для уровня значимости 0,05.

Таблица 2.8

i	Интервалы зарплат	Частоты по отрасли А	Частоты по отрасли В	Частоты для объединения выборок	Относительные частоты для объединения выборок
		j=1	j=2		
1	130-150	4	1	5	5/200
2	150-170	4	1	5	5/200
3	170-200	15	8	23	23/200
4	200-250	51	43	94	94/200
5	250-300	22	34	56	56/200
6	300-350	3	7	10	10/200
7	350-400	1	3	4	4/200
8	400-450	0	3	3	3/200
	Σ	100	100	200	

Вычисление значения критерия χ^2 :

$$\chi^2 = \left[\frac{\left(4 - 100 \cdot \frac{5}{200}\right)^2}{100 \cdot \frac{5}{200}} + \frac{\left(4 - 100 \cdot \frac{5}{200}\right)^2}{100 \cdot \frac{5}{200}} + \dots + \frac{\left(0 - 100 \cdot \frac{3}{200}\right)^2}{100 \cdot \frac{3}{200}} \right] + \left[\frac{\left(1 - 100 \cdot \frac{5}{200}\right)^2}{100 \cdot \frac{5}{200}} + \frac{\left(1 - 100 \cdot \frac{5}{200}\right)^2}{100 \cdot \frac{5}{200}} + \dots + \frac{\left(3 - 100 \cdot \frac{3}{200}\right)^2}{100 \cdot \frac{3}{200}} \right]$$

или $\chi^2 = 200 / 100 \cdot [(4 - 2,5)^2 / 5 + (4 - 2,5)^2 / 5 + \dots + (0 - 1,5)^2 / 3 + (1 - 2,5)^2 / 5 + (1 - 2,5)^2 / 5 + \dots + (3 - 1,5)^2 / 3] = 14,58$,

$k=2$, $m=8$, следовательно, число степеней свободы равно 7 и $\chi'_{\text{крит}} = 14,067$.

Таким образом, $\chi^2_{\text{набл}} > \chi^2_{\text{крит}}$ и на уровне значимости 0,05 гипотеза о том, что распределение зарплат в отраслях А и В совпадает, не проходит.

2.7. Критерий независимости

Похожая схема применяется для проверки зависимости признаков А и В. Пусть имеются два признака, которые мы условно обозначим через А и В (например, количество внесенных удобрений и урожайность, размер инвестиций и размер прибыли и т.д.). При исследовании измеряются значения пары (а,в). По обоим признакам производится группировка. Пусть по признаку А число интервалов группировки равно r , а по признаку В – s . Результаты выборки сводятся в двумерную таблицу (табл. 2.9).

В табл. 2.9 через n_{ij} обозначено количество элементов выборки, у которых значение признака А попало в i -й интервал признака А и вместе с тем значение признака В попало в j -й интервал признака В.

Если признаки А и В независимы друг от друга и численность выборки n большая, то величина

$$\chi^2 = n \sum_{i=1}^r \sum_{j=1}^s \frac{\left(n_{ij} - \frac{a_i b_j}{n}\right)^2}{a_i b_j}$$

Таблица 2.9

Интервалы значений	B_1	B_2	...	B_s	Сумма
A_1	n_{11}	n_{12}	...	n_{1s}	a_1
A_2	n_{21}	n_{22}	...	n_{2s}	a_2
...	...	...	...	...	...
A_r	n_{r1}	n_{r2}	...	n_{rs}	a_r
Сумма	b_1	b_2	...	b_s	n

имеет приближенно χ^2 -распределение с $(r - 1)(s - 1)$ степенями свободы. Критическая область для проверки гипотезы, что признаки A и B независимы друг от друга, определена на уровне значимости α неравенством $\chi^2 > \chi^2_\alpha$ при $f = (r - 1)(s - 1)$ степенях свободы. Вопрос о зависимости признаков мы в дальнейшем рассмотрим подробнее.

3. ДИСПЕРСИОННЫЙ АНАЛИЗ. СРАВНЕНИЕ ПАРАМЕТРОВ БОЛЕЕ ЧЕМ ДВУХ СЛУЧАЙНЫХ ВЕЛИЧИН

Нами рассматривалась задача проверки значимости различия выборочных средних двух совокупностей. На практике часто возникает необходимость обобщения задачи, т.е. проверки существенности различия выборочных средних m совокупностей ($m > 2$). Например, требуется оценить влияние различных свойств сырья на показатели качества продукции, количества вносимых удобрений на урожайность и т. д. Для эффективного решения такой задачи нужен новый подход, который и реализуется в дисперсионном анализе.

Дисперсионный анализ – это статистический метод, предназначенный для оценки влияния различных факторов на результат эксперимента:

Дисперсионный анализ был разработан английским математиком Р.А.Фишером в 1918 году, хотя сам термин “дисперсионный анализ” он употребил позднее.

По числу факторов, влияние которых изучается, различают однофакторный и многофакторный анализ. Мы будем заниматься подробно только **однофакторным анализом**, когда исследуется наличие или отсутствие влияния на признак одного фактора. Для изучения результирующего признака для каждого уровня фактора проводится n испыта-

ний. Обозначим через x_{ij} значение измеряемого признака X , полученное в i -м опыте при j -м уровне фактора. Под уровнем фактора понимается его мера или состояние, например номер партии деталей, количество вносимых удобрений и т.д. Если фактор, влияние которого изучается, принимает числовое значение – это номер значения или номер интервала группировки.

Например, экономиста строительного-монтажного управления интересует зависимость объема выполненных на стройке работ за смену от работающей на стройке бригады. Предположим, что на стройке работает m бригад. Изучаемым результативным признаком является объем выполненных работ. Его обозначим через X . Работающую бригаду назовем фактором A , а номер работающей бригады – уровнем или группой фактора A и через A_j будем обозначать j -й уровень или j -ю группу фактора A (j -ю бригаду, $j=1,2,\dots,m$). Над каждой из m бригад n раз проводится наблюдение и результаты наблюдений заносятся в табл. 3.1.

Таблица 3.1

$i \backslash j$	1	2	...	m
Ном. фактор	A_1	A_2	...	A_m
Ном. набл.				
1	x_{11}	x_{12}	...	x_{1m}
2	x_{21}	x_{22}	...	x_{2m}
...	...	...	...	...
n	x_{n1}	x_{n2}	...	x_{nm}
Выборочное среднее	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_m$
Выборочная дисперсия	S^2_1	S^2_2	...	S^2_m

Здесь $\bar{x}_j = 1/n(x_{1j} + x_{2j} + \dots + x_{nj})$ и $S^2_j = 1/n \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$ (выборочное среднее и выборочная дисперсия на j -м уровне фактора).

Такую же таблицу заполнит исследователь, изучающий существен-

ность влияния партий изделий на их качество, зависимость урожайности от номера участка и т.д.

С точки зрения теории вероятностей случайная величина X_{ij} в схеме однофакторного анализа есть сумма двух компонентов:

$$X_{ij} = a_j + e_{ij},$$

где a_j – математическое ожидание признака (объема выполненных работ за смену), а e_{ij} – случайное отклонение признака от математического ожидания (возмущение) в i -м испытании при j -м уровне фактора.

В схеме дисперсионного анализа предполагается, что возмущения e_{ij} взаимно независимы, имеют нормальное распределение $N(0, \sigma)$ (нулевое математическое ожидание и одинаковую дисперсию σ^2). Таким образом, задача сводится к проверке нулевой гипотезы H_0 : $a_1 = a_2 = \dots = a_m$, когда имеется m выборок объема n из m нормальных законов с одинаковыми дисперсиями σ^2 .

Обозначим через $\bar{x}$ общую среднюю

$$\bar{x} = \frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n x_{ij} = \frac{1}{m} \sum_{i=1}^m \bar{x}_i$$

и рассмотрим сумму квадратов отклонений наблюдений x_{ij} от общей средней $\bar{x}$

$$Q_{\text{общ}} = \sum_{j=1}^m \sum_{i=1}^n (x_{ij} - \bar{x})^2.$$

Путем простых выкладок можно показать, что

$$Q_{\text{общ}} = Q_{\text{факт}} + Q_{\text{ост}},$$

где $Q_{\text{факт}} = n \sum_{j=1}^m (\bar{x}_j - \bar{x})^2$ – факторная (межгрупповая) сумма квадратов отклонений групповых средних от общей средней, а

$$Q_{\text{ост}} = \sum_{j=1}^m \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$$
 – остаточная (внутригрупповая) сумма квадратов отклонений от групповых средних.

Для простоты в дальнейшем будем обозначать

$$Q_{\text{общ}} = Q; Q_{\text{факт}} = Q_1; Q_{\text{ост}} = Q_2,$$

т.е. $Q = Q_1 + Q_2$.

Последнее равенство показывает, что общая вариация признака складывается из двух компонентов: вариации изменчивости этого показателя между "партиями" и вариации внутри "партий", характеризующей одинаковую для всех партий вариацию под воздействием неучтенных факторов (по условию дисперсии случайных вариаций одинаковы и равны σ^2).

В дисперсионном анализе анализируются "средние квадраты" – суммы квадратов отклонений, деленные на соответствующее число степеней свободы. Они являются несмещенными оценками соответствующих дисперсий. Число степеней свободы – это число квадратов в сумме минус число связывающих их уравнений. Поэтому для s_1^2 число степеней свободы равно $m - 1$, а для s_2^2 составляет $mn - m$.

Таким образом, $s_1^2 = Q_1 / (m - 1)$, а $s_2^2 = Q_2 / (mn - m)$. Можно показать, что s_1^2 и s_2^2 независимы друг от друга, т.е. s_1^2 и s_2^2 являются несмещенными и независимыми оценками одной и той же дисперсии σ^2 .

Следовательно, проверка нулевой гипотезы о равенстве m математических ожиданий свелась к проверке существенности различия несмещенных выборочных оценок s_1^2 и s_2^2 . Она осуществляется с по-

мощью статистики $F = \frac{s_1^2}{s_2^2}$.

Гипотеза H_0 отвергается, если фактически вычисленное значение статистики $F > F_{\alpha; k_1; k_2}$, определенного на уровне значимости α при числе степеней свободы $k_1 = m - 1$, $k_2 = mn - m$, и принимается, если $F < F_{\alpha; k_1; k_2}$. Опровержение гипотезы означает наличие существенных различий в зависимости от уровня фактора.

Исследуемую модель, которую мы описали как $x_{ij} = a_j + e_{ij}$, где a_j – среднее значение для j -го уровня фактора, а e_{ij} – случайное отклонение от среднего, можно представить и в такой форме:

$$x_{ij} = a + \lambda_j + e_{ij},$$

где a – генеральное среднее всех мыслимых результатов наблюдений или, иначе говоря, математическое ожидание результативного признака X , а λ_j – генеральный эффект влияния на X , вызванный j -м уровнем фактора, или, иначе, отклонение математического ожидания a_j результативного признака при j -м уровне фактора от общего математического ожидания a , т.е. $\lambda_j = a_j - a$. Можно сказать, что мы изучаем влияние фактора A для параметрической модели, параметрами которой являются математическое ожидание a , эффекты $\lambda_1, \lambda_2, \dots, \lambda_m$ влияния уровней фак-

тора и дисперсия σ^2 случайного остатка. Если гипотеза H_0 о равенстве математических ожиданий при разных уровнях фактора не отвергается, т.е. влияние фактора A не подтверждается, то эффекты влияния уровней фактора λ_j будут нулевыми, остальные параметры по результатам наблюдений оцениваются следующим образом.

Оценка $\bar{x}$ равна общему среднему по всем наблюдениям

$$\bar{x} = \frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n x_{ij}.$$

Оценка для σ^2 :

$$s_{\text{ост}}^2 = \frac{1}{mn - m} \sum_{j=1}^m \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2.$$

Если же гипотеза H_0 отвергается, то есть влияние фактора подтверждается по результатам наблюдений, то оценки для λ_j уже будут не нулевыми, а равными разности $\bar{x}_j - \bar{x}$, где $\bar{x}_j$ – групповое среднее на j-м уровне фактора A.

Рассмотрим конкретный пример о зависимости объема работ, выполненных за смену, от работающей бригады. Фактором будет номер бригады ($j=1,2,3,4$), для каждой бригады произведено четыре наблюдения (табл. 3.2).

Таблица 3.2

$i \setminus j$	1	2	3	4
Ном. набл. \ фактор	A_1	A_2	A_3	A_4
1	140	150	148	150
2	144	149	149	155
3	142	152	146	154
4	145	150	147	152
Выборочное среднее	$\bar{x}_1=142,75$	$\bar{x}_2=150,25$	$\bar{x}_3=147,5$	$\bar{x}_4=152,75$
Выборочная дисперсия	$S^2_1=3,69$	$S^2_2=1,19$	$S^2_3=1,25$	$S^2_4=3,69$

Проверим по F-критерию значимости влияние номера бригады на выработку за смену.

1. Вычислим средний объем работ по всем 16-ти наблюдениям:

$$\bar{x} = 1/16(140+144+142+145+150+\dots+154+152) = 148,31.$$

2. Вычислим факторную и остаточную вариации и оценки для дисперсий, участвующих в F-критерии:

$$Q_{\text{факт}} = n \sum_{j=1}^m (\bar{x}_j - \bar{x})^2 = 4[(142,75-148,31)^2 + (150,25-148,31)^2 + (147,5-148,31)^2 + (152,75-148,31)^2] = 220,19;$$

$$Q_{\text{ост}} = \sum_{j=1}^m \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 = 3,69 \cdot 4 + 1,19 \cdot 4 + 1,25 \cdot 4 + 33,69 \cdot 4 = 39,27;$$

$$Q_{\text{общ}} = 220,19 + 39,27 = 259,46.$$

Составим дисперсионную таблицу (табл. 3.3).

Таблица 3.3

Источник вариации объема работ	Сумма квадратов отклонений	Число степеней свободы	Несмещенная оценка для дисперсии
Работающая бригада	220,19	4-1=3	$s^2_{\text{факт}} = 220,19/3 = 73,4$
Остаточные факторы	39,27	16-4=12	$s^2_{\text{ост}} = 39,17/12 = 3,27$
Общая вариация	259,16	16-1	$259,46/15 = 17,30$

Теперь можно вычислить наблюдаемое значение F-критерия:

$$F_{\text{набл}} = s^2_{\text{факт}} / s^2_{\text{ост}} = 73,4/3,27 = 22,45;$$

$$F_{\text{крит}} = F_{0,05;3;12} = 3,49.$$

Так как $F_{\text{набл}} > F_{\text{крит}}$, то гипотезу H_0 отвергаем, т.е. считаем, что объем ежедневной выработки у разных бригад разный.

Определим **коэффициент детерминации** как отношение факторной вариации к общей (т.е. коэффициент детерминации указывает,

какая доля дисперсии признака X объясняется зависимостью признака X от фактора):

$$\rho^2 = Q_{\text{факт}} / Q_{\text{общ}}.$$

В нашем конкретном случае

$$\rho^2 = 220,19 / 259,16 = 0,849.$$

Это означает, что 84,9% общей вариации ежедневного объема выработки связано с работающей бригадой.

Оценка средней выработки за смену по совокупности бригад (на стройке в целом) равна 148,31.

Оценки “эффекта бригады” соответственно равны:

$$\lambda_1 = 142,75 - 148,31 = -5,56;$$

$$\lambda_2 = 150,25 - 148,31 = 1,94;$$

$$\lambda_3 = 147,5 - 148,31 = -0,81;$$

$$\lambda_4 = 152,75 - 148,31 = 4,44.$$

Знак λ_j , а также ее величина являются характерными показателями уровня фактора (конкретной бригады).

Из свойств математического ожидания и дисперсии следует, что для упрощения расчетов из всех значений x_{ij} можно вычесть одну и ту же константу С. Кроме того, все значения x_{ij} можно умножить на один и тот же множитель (так как статистика F и коэффициент детерминации есть отношение двух дисперсий). Так что если наблюдаемые значения заданы десятичными числами с одним знаком после запятой, то их целесообразно умножить на 10, если с двумя – на 100 и т.д. Если задано k знаков после запятой, то для того чтобы в качестве выборки фигурировали сравнительно небольшие целые числа, целесообразно вместо x_{ij} взять $y_{ij} = 10^k x_{ij} - C$. Например, если $x_{11} = 12,1$, $x_{12} = 12,2$, $x_{31} = 12,6$, то, приняв $y_{ij} = 10x_{ij} - 123$, получим: $y_{11} = 121 - 123 = -2$, $y_{21} = 122 - 123 = -1$, $y_{31} = 126 - 123 = 3$.

Замечание 1. Если факторная дисперсия окажется меньше остаточной [т.е. вариация, вызванная фактором, меньше вариации, вызванной остальными (случайными) причинами], то уже отсюда следует справедливость гипотезы о равенстве групповых средних и, значит, нет надобности прибегать к критерию F.

Замечание 2. Если нет уверенности в справедливости предположения о равенстве дисперсий рассматриваемых совокупностей, то это предположение следует предварительно проверить.

В предыдущем разделе мы приводили критерий проверки равенства двух дисперсий (с помощью статистики F). Проверка гипотезы о

равенстве m дисперсий осуществляется с помощью **критерия Бартлета**, использующего статистику, имеющую распределение χ^2 .

Проверка гипотезы о равенстве m дисперсий по критерию Бартлета выполняется по следующему алгоритму.

Пусть имеются m нормально распределенных совокупностей, дисперсии которых равны соответственно $\sigma_1^2, \sigma_2^2, \dots, \sigma_m^2$, и m независимых выборок из каждой совокупности объема $n_1, n_2, \dots, n_m$. Необходимо проверить нулевую гипотезу о равенстве дисперсий $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_m^2 = \sigma^2$.

Вычислим для всех j выборок несмещенные оценки s_j^2 для дисперсий σ_j^2 и найдем величину s_0^2 по формуле

$$s_0^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_m - 1)s_m^2}{(n_1 - 1) + (n_2 - 1) + \dots + (n_m - 1)}.$$

Далее найдем

$$q = \left[1 + \frac{1}{3(m-1)} \left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} + \dots + \frac{1}{n_m - 1} + \frac{1}{(n_1 - 1) + (n_2 - 1) + \dots + (n_m - 1)} \right) \right]^{-1}.$$

Доказано, что при справедливости гипотезы H_0 , если при всех j выполняется $n_j > 3$, то статистика Φ

$$\Phi = q \left[(n_1 - 1) \ln \frac{s_0^2}{s_1^2} + (n_2 - 1) \ln \frac{s_0^2}{s_2^2} + \dots + (n_m - 1) \ln \frac{s_0^2}{s_m^2} \right],$$

имеет распределение, близкое к χ^2 -распределению с $m - 1$ степенями свободы. Следовательно, для проверки гипотезы H_0 надо найти в табл. 2 приложения 2 для заданного уровня значимости α соответствующее критическое значение $\chi^2_{\alpha; m-1}$ и сравнить с ним величину Φ . Если окажется, что $\Phi < \chi^2_{\alpha; m-1}$, то можно сказать, что на уровне значимости α гипотеза о равенстве дисперсий для m совокупностей подтверждается.

Посмотрим, что дала бы проверка по критерию Бартлета гипотезы о равенстве дисперсий для наших строительных бригад:

$$s_1^2 = S_1^2 \cdot n_1 / (n_1 - 1) = 3,69 \cdot 4 / 3 = 4,92; \quad s_2^2 = S_2^2 \cdot n_2 / (n_2 - 1) = 1,19 \cdot 4 / 3 = 1,58;$$

$$s_3^2 = 1,67; \quad s_4^2 = 4,92;$$

$$s_0^2 = 1 / (3 + 3 + 3 + 3) (3 \cdot 4,92 + 3 \cdot 1,58 + 3 \cdot 1,67 + 3 \cdot 4,92) = 3,27;$$

$$q = [1 + 1/3(4-1)(1/3 + 1/3 + 1/3 + 1/3 + 1/(3+3+3+3))]^{-1} = 0,878;$$

$$\Phi = 0,878 \cdot 3 [\ln 3,27 / 4,92 + \ln 3,27 / 1,58 + \ln 3,27 / 1,67 + \ln 3,27 / 4,92] = 1,538;$$

$$\chi^2_{\alpha; m-1} = \chi^2_{0,05; 4-1} = 7,82.$$

Так как $1,538 < 7 < 7,82$, то гипотеза о равенстве дисперсий для всех четырех бригад на уровне значимости 0,05 проходит.

Замечание. Если гипотеза о равенстве дисперсий принимается, за оценку общего значения дисперсии принимают величину s_0^2 , вычисляемую по формуле:

$$s_0^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_m - 1)s_m^2}{(n_1 - 1) + (n_2 - 1) + \dots + (n_m - 1)}.$$

Заметим, что часто для расчетов удобно пользоваться формулами, не требующими вычисления самих средних:

$$Q_1 = \frac{1}{n} \sum_{j=1}^m \left(\sum_{i=1}^n x_{ij} \right)^2 - \frac{1}{mn} \left(\sum_{j=1}^m \sum_{i=1}^n x_{ij} \right)^2;$$

$$Q_2 = \sum_{j=1}^m \sum_{i=1}^n x_{ij}^2 - \frac{1}{n} \sum_{j=1}^m \left(\sum_{i=1}^n x_{ij} \right)^2;$$

$$Q = \sum_{j=1}^m \sum_{i=1}^n x_{ij}^2 - \frac{1}{mn} \left(\sum_{j=1}^m \sum_{i=1}^n x_{ij} \right)^2.$$

Введем обозначения:

$R_j = \sum_{i=1}^n x_{ij}^2$; $T_j = \sum_{i=1}^n x_{ij}$ или $R_j = \sum_{i=1}^n y_{ij}^2$; $T_j = \sum_{i=1}^n y_{ij}$, если мы перешли от x_{ij} к $y_{ij} = 10^k x_{ij} - C$.

Тогда:

$$Q = \sum_{j=1}^m \sum_{i=1}^n x_{ij}^2 - \frac{1}{mn} \left(\sum_{j=1}^m \sum_{i=1}^n x_{ij} \right)^2 = \sum_{j=1}^m R_j - \frac{1}{mn} \left(\sum_{j=1}^m T_j \right)^2;$$

$$Q_1 = \frac{1}{n} \sum_{j=1}^m \left(\sum_{i=1}^n x_{ij} \right)^2 - \frac{1}{mn} \left(\sum_{j=1}^m \sum_{i=1}^n x_{ij} \right)^2 = \frac{1}{n} \sum_{j=1}^m T_j^2 - \frac{1}{mn} \left(\sum_{j=1}^m T_j \right)^2;$$

$$Q_{\text{ост}} = Q_{\text{общ}} - Q_{\text{факт}} (Q_2 = Q - Q_1).$$

Формулы легко модифицируются для случая, когда разным уровням фактора соответствует разное число наблюдений (см., например, работу [1]).

Пример 3.1. Произведено четыре испытания на каждом из трех уровней. Результаты испытаний приведены в табл. 3.4.

Таблица 3.4

i	F ₁	F ₂	F ₃
1	51	52	42
2	52	54	44
3	56	56	50
4	57	58	52
$\bar{x}_j$	54	55	47

Методом дисперсионного анализа на уровне значимости 0,05 проверить гипотезу о равенстве средних, если выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями.

Решение. Для упрощения расчета вычтем $C=52$ из каждого наблюдаемого значения ($52=1/3(54+55+47)$) и составим расчетную табл. 3.5.

Таблица 3.5

i\j	F ₁		F ₂		F ₃		Σ
	y_{i1}	y_{i1}^2	y_{i2}	y_{i2}^2	y_{i3}	y_{i3}^2	
1	-1	1	0	0	-10	100	
2	0	0	2	4	-8	64	
3	4	16	4	16	-2	4	
4	5	25	6	36	0	0	
$R_j = \sum_{i=1}^n y_{ij}^2$		42		56		168	$\Sigma R_j = 266$
$T_j = \sum_{i=1}^n y_{ij}$	8		12		-20		$\Sigma T_j = 0$
T_j^2	64		144		400		$\Sigma T_j^2 = 608$

$$Q = \sum_{j=1}^m \sum_{i=1}^n x_{ij}^2 - \frac{1}{mn} \left(\sum_{j=1}^m \sum_{i=1}^n x_{ij} \right)^2 = \sum_{j=1}^m R_j - \frac{1}{mn} \left(\sum_{j=1}^m T_j \right)^2 = 266 - 0 = 266;$$

$$Q_1 = \frac{1}{n} \sum_{j=1}^m \left(\sum_{i=1}^n x_{ij} \right)^2 - \frac{1}{mn} \left(\sum_{j=1}^m \sum_{i=1}^n x_{ij} \right)^2 = \frac{1}{n} \sum_{j=1}^m T_j^2 - \frac{1}{mn} \left(\sum_{j=1}^m T_j \right)^2 = (608/4) - 0 = 152;$$

$$Q_2 = Q_{\text{ост}} = Q - Q_1 = Q_{\text{общ}} - Q_{\text{факт}} = 266 - 152 = 114.$$

Составим дисперсионную таблицу 3.6.

Таблица 3.6

Источник вариации	Сумма квадратов отклонений	Число степеней свободы	Несмещенная оценка для дисперсии
Влияние изучаемого фактора	152	3-1=2	$s^2_{\text{факт}}=152/2=76$
Остаточные факторы	114	12-3=9	$s^2_{\text{ост}}=114/9=12,67$
Общая вариация	266	12-1=11	$266/11=24,2$

Найдем факторную и остаточную дисперсии:

$$s^2_{\text{факт}} = Q_{\text{факт}} / (m-1) = 152 / (3-1) = 76;$$

$$s^2_{\text{ост}} = Q_{\text{ост}} / n(m-1) = 114 / 3(4-1) = 12,67.$$

Сравним факторную и остаточную дисперсии по критерию F, для чего найдем наблюдаемое значение критерия:

$$F_{\text{набл}} = s^2_{\text{факт}} / s^2_{\text{ост}} = 76 / 12,67 = 6.$$

Учитывая, что число степеней свободы числителя $k_1 = 3 - 1 = 2$, а знаменателя $k_2 = 3 \cdot (4 - 1) = 9$ и уровень значимости $\alpha = 0,05$, по табл. 3 критических точек распределения F находим критическую точку:

$$F_{\text{кр}}(0,05; 2; 9) = 4,26.$$

Так как $F_{\text{набл}} > F_{\text{кр}}$, то нулевую гипотезу о равенстве средних при разных уровнях фактора отвергаем. Средние "в целом" различаются значительно.

Оценка генерального среднего для всех уровней фактора равна $0+52=52$.

Оценки "эффекта уровня фактора" соответственно равны:

$$\lambda_1=54-52=2;$$

$$\lambda_2=55-52=1;$$

$$\lambda_3=47-52=-5.$$

Знак λ_1 , а также и ее величина являются характерными показателями уровня фактора. Первые два уровня мало различаются. Последний уровень имеет сильное отрицательное влияние.

Коэффициент детерминации:

$$\rho^2=Q_{\text{факт}}/Q_{\text{общ}}=152/266=0,57.$$

Если требуется средние сравнить попарно, то следует воспользоваться критерием Стьюдента. Например, по Стьюденту можно проверить значимость различия первого и второго уровня фактора.

Описание исследования многофакторных моделей, когда на результат влияет несколько факторов, можно найти, например, в работе [2]. Коротко остановимся на схеме двухфакторного анализа.

Пусть изучается влияние на результирующий признак X двух факторов – A и B . Все имеющиеся данные представим в виде таблицы, в которой по горизонтали – уровни фактора A , по вертикали – уровни фактора B , а в соответствующих клетках, или ячейках, таблицы находятся показатели исследуемого признака для данных уровней фактора. Как таблица, так и формулы зависят от количества проведенных наблюдений при каждой комбинации A_i, B_j уровней факторов или, иначе говоря, в каждой клетке двухфакторного комплекса. В случае, если в каждой клетке проведено одно наблюдение, таблица наблюдений имеет следующий вид (табл. 3.7).

Таблица 3.7

$A \setminus B$		B_1	B_2	...	B_k
	$i \setminus j$	1	2	...	k
A_1	1	x_{11}	x_{12}		x_{1k}
A_2	2	x_{21}	x_{22}		x_{2k}
...	...				
A_m	m	x_{m1}	x_{m2}		x_{mk}

Если в каждой клетке проведено больше, чем одно наблюдение, таблица приобретает более сложный вид. Приведем пример таблицы, когда для каждой комбинации A_i , B_j уровней факторов проводилось по три испытания (табл. 3.8).

Таблица 3.8

B\A	A ₁			A ₂			A ₃		
Номер наблюдения	1	2	3	1	2	3	1	2	3
B ₁	50	54	58	62	60	58	65	71	65
B ₂	54	46	50	64	59	60	59	54	61
B ₃	52	48	50	70	62	60	59	66	64
B ₄	60	55	56	58	54	50	71	74	62

Например, такую таблицу составили, изучая влияние на производительность труда трех технологий производства однотипных деталей на четырех заводах. Производительность труда измеряли в некоторых условных единицах.

Двухфакторная дисперсионная модель принимает вид

$$x_{ijk} = a + F_i + G_j + I_{ij} + e_{ijk},$$

где F_i – эффект, обусловленный влиянием i -го уровня фактора A ;
 G_j – эффект, обусловленный влиянием j -го уровня фактора B ;
 I_{ij} – эффект, обусловленный взаимодействием двух факторов для i -го уровня фактора A и j -го уровня фактора B ;
 e_{ijk} – возмущение, обусловленное вариацией переменной внутри отдельной ячейки.

Все теоретические рассуждения для двухфакторной модели абсолютно аналогичны тем, что мы проводили для однофакторной модели. Но формулы, естественно, становятся более громоздкими. При решении реальных задач методом дисперсионного анализа обычно используют статистические программные пакеты.

При неравном числе наблюдений в ячейках резко возрастает сложность аппарата дисперсионного анализа, поэтому рекомендуется планировать схему с равным числом наблюдений в ячейках.

4. КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ. ВЫЯВЛЕНИЕ СВЯЗИ МЕЖДУ ВЕЛИЧИНАМИ

Во многих задачах требуется установить и оценить зависимость случайной величины Y от случайной или неслучайной величины X (или от нескольких величин).

Две случайные величины могут быть независимыми, если они связаны либо функциональной зависимостью (значение Y однозначно вычисляется по значению X), либо статистической зависимостью, когда кроме физической зависимости обе величины или одна из них подвержены влиянию случайных факторов.

Статистической называют зависимость, при которой изменение одной величины влечет изменение распределения другой. Самый простой случай – при изменении одной величины изменяется среднее значение (математическое ожидание) другой. В этом случае статистическую зависимость называют **корреляционной**.

Например, пусть Y – урожай зерна, X – количество удобрений. С одинаковых по площади участков земли при равных количествах внесенных удобрений снимают различный урожай, т.е. Y не является функцией X . Это объясняется влиянием различных случайных факторов (осадки, уход, тип почвы и др.). Вместе с тем, как показывает опыт, средний урожай является функцией от количества удобрений, т.е. Y связан с X корреляционной зависимостью.

В терминах теории вероятностей при статистической зависимости величины Y и X как случайные величины задаются своим совместным распределением вероятностей.

4.1. Двумерные дискретные случайные величины.

Понятие условного распределения.

Условные математическое ожидание и дисперсия

Напомним, что распределение двумерной дискретной случайной величины (X, Y) задается с помощью двумерной таблицы (табл. 4.1).

В этой таблице $p_{ij} = P(X=x_i, Y=y_j)$, а с помощью символа $*$ указывает, что произведено суммирование по индексу, на месте которого стоит $*$. Можно сказать, что значения $p_{.j}$ ($j=1, \dots, k$) задают распределение случайной величины X [$p_{.j}$ – это вероятность того, что $P(X=x_i)$], а значения $p_{i.}$ ($i=1, \dots, m$) задают распределение случайной величины Y [$p_{i.} = P(Y=y_j)$]. Для этих распределений можно вычислить математическое ожидание (среднее значение) и дисперсию:

Таблица 4.1

$Y \backslash X$	x_1	x_2	...	x_k	$\sum_{j=1}^k$
y_1	$p_{11}=P(x_1, y_1)$	$p_{12}=P(x_2, y_1)$	...	$p_{1k}=P(x_k, y_1)$	$p_{1*}=\sum p_{1j}$
y_2	$p_{21}=P(x_1, y_2)$	p_{22}	...	p_{2k}	$p_{2*}=\sum p_{2j}$
...	...	...	...	...	...
y_m	p_{m1}	p_{m2}	...	p_{mk}	$p_{m*}=\sum p_{mj}$
$\sum_{i=1}^m$	$p_{*1}=\sum p_{i1}$	$p_{*2}=\sum p_{i2}$	...	$p_{*k}=\sum p_{ik}$	1

$$M(X)=\sum x_i p_{i*}, \quad D(X)=\sum (x_i - M(X))^2 p_{i*};$$

$$M(Y)=\sum y_i p_{i*}, \quad D(X)=\sum (y_i - M(X))^2 p_{i*}.$$

Зафиксировав значение $x=x_i$, можно получить условное распределение $P(Y/X=x_i)$:

$$P\{Y=y_i/X=x_j\} = P(X=x_j, Y=y_i) / P(X=x_j) = p_{ij}/p_{j*}.$$

Аналогично находим условное распределение $P(X/Y=y_i)$:

$$P(X=x_j/Y=y_i) = P(X=x_j, Y=y_i) / P(Y=y_i) = p_{ij}/p_{i*}.$$

Условные распределения – это распределения вероятностей. По ним можно сосчитать условные математические ожидания $M(Y/X=x)$ и $M(X/Y=y)$ и условные дисперсии $D(Y/X=x)$ и $D(X/Y=y)$. Они вычисляются по формулам:

$$M(Y/X=x_j) = \sum_{i=1}^m y_i P(Y=y_i/X=x_j) = 1/p_{j*} \sum_{i=1}^m y_i p_{ij};$$

$$D(Y/X=x_j) = \sum_{i=1}^m (y_i - M(Y/X=x_j))^2 P(Y=y_i/X=x_j);$$

$$M(X/Y=y_i) = \sum_{j=1}^k x_j P(X=x_j/Y=y_i) = 1/p_{i*} \sum_{j=1}^k x_j p_{ij};$$

$$D(X/Y=y_i) = \sum_{j=1}^k (x_j - M(X/Y=y_i))^2 P(X=x_j/Y=y_i).$$

Пример 4.1. Распределение двумерной дискретной случайной величины задано с помощью табл. 4.2.

Таблица 4.2

$Y \backslash X$	$x_1=1$	$x_2=3$	$x_3=4$	$x_4=8$	Σ
$y_1=3$	0,15	0,06	0,25	0,04	0,5
$y_2=6$	0,30	0,10	0,03	0,07	0,5
Σ	0,45	0,16	0,28	0,11	1

Найдем условное распределение величины Y при условии $X=x_1=1$ и условное математическое ожидание Y при условии $X=x_1=1$:

$$P(y_1/x_1)=P(x_1, y_1)/P(x_1)=0,15/0,45=1/3;$$

$$P(y_2/x_1)=P(x_1, y_2)/P(x_1)=0,30/0,45=2/3;$$

$$M(Y/X=1)=1/0,45 \cdot (3 \cdot 0,15 + 6 \cdot 0,30) = 3 \cdot 1/3 + 6 \cdot 2/3 = 1 + 4 = 5.$$

Аналогично находим условные математические ожидания Y при остальных значениях X и условные математические ожидания X при условии Y :

$$M(Y/X=3)=(3 \cdot 0,06 + 6 \cdot 0,10)/0,16=4,875;$$

$$M(Y/X=4)=(3 \cdot 0,25 + 6 \cdot 0,03)/0,28=3,32...;$$

$$M(Y/X=8)=(3 \cdot 0,04 + 6 \cdot 0,07)/0,11=4,91...;$$

$$M(X/Y=3)=(1 \cdot 0,15 + 3 \cdot 0,06 + 4 \cdot 0,25 + 8 \cdot 0,04)/0,5=3,3;$$

$$M(X/Y=6)=(1 \cdot 0,30 + 3 \cdot 0,10 + 4 \cdot 0,03 + 8 \cdot 0,07)/0,5=2,56.$$

4.2. Регрессия, линейная регрессия. Коэффициент корреляции

Условное математическое ожидание $M(Y/X=x)$ есть функция от x :

$$M(Y/X=x)=f(x).$$

Эту функцию называют **функцией регрессии Y на X** .

Вообще говоря, когда изучают пары величин (X, Y) , в которых одна величина неслучайна, например время или координата в пространстве, то говорят о **регрессионной модели**.

В случае, когда функция регрессии – линейная функция, т.е. ее

можно представить в виде уравнения $y=f(x)=\alpha x+\beta$, **регрессия** называется **линейной**.

Если в паре (X,Y) обе величины случайны и равноправны, то **модель** называют **ковариационной** (covariation) или **корреляционной** (correlation – взаимозависимость).

В ковариационной модели определяются две функции регрессии. Кроме регрессии Y на X , определяется и **функция регрессии X на Y** :

$$M(X/Y=y)=\varphi(y).$$

Пример 4.2. Изобразить на координатной плоскости сосчитанные в предыдущем примере условные математические ожидания (рис.4.1).

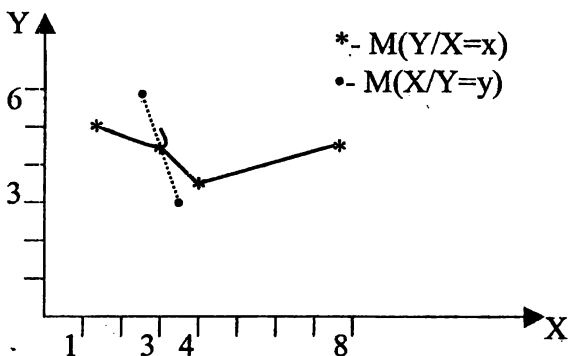


Рис. 4.1

Если обе функции регрессии – линейные (т.е. $y=f(x)=\alpha x+\beta$ и $x=\varphi(y)=\Theta y+\Delta$), то говорят, что X и Y связаны **линейной корреляционной зависимостью**.

Линейность функции регрессии есть показатель **вида связи**. Регрессия может выражаться с помощью функции другого вида, например параболы. Функциональная зависимость условного математического ожидания от значения второй компоненты является одним из показателей зависимости пары величин. Но это не единственный показатель зависимости величин. Даже если условные математические ожидания с изменением второй компоненты не меняются, могут меняться условные дисперсии или другие показатели.

Кроме вида связи интересно оценить **тесноту** зависимости двух величин. Важным **числовым** показателем зависимости пары случай-

ных величин является значение **ковариации** их совместного распределения, а точнее, их **коэффициента корреляции**.

Ковариация двух случайных величин определяется как математическое ожидание их произведения минус произведение их математических ожиданий:

$$K_{xy} = M[(X - M(X))(Y - M(Y))] = M(XY) - M(X)M(Y).$$

Для двумерного дискретного распределения ковариация вычисляется по следующей формуле:

$$K_{xy} = \sum \sum x_i y_j p_{ij} - \sum x_i p_{i\cdot} \cdot \sum y_j p_{\cdot j}.$$

Ковариация – величина размерная, поэтому вместо нее обычно употребляют **коэффициент корреляции** – отношение ковариации к произведению среднеквадратических отклонений величин X и Y:

$$\sqrt{\rho_{xy}} = K_{xy} / \sigma_x \sigma_y.$$

Коэффициент корреляции обладает следующими свойствами:

1. $-1 \leq \rho \leq 1$.

2. Если случайные величины независимы, то их коэффициент корреляции равен нулю. Обратное утверждение, вообще говоря, неверно: из того, что $\rho = 0$, еще не следует независимость случайных величин.

3. Если у двух случайных величин $|\rho| = 1$, то между этими случайными величинами существует линейная функциональная зависимость; причем если связь по возрастающей (большему значению X соответствует большее значение Y), то $\rho = 1$, а если связь по убывающей (большему значению X соответствует меньшее значение Y), то $\rho = -1$.

Это вытекает из того, что

$$\begin{aligned} M \left(\frac{X - M(X)}{\sigma_x} \pm \frac{Y - M(Y)}{\sigma_y} \right)^2 &= \\ &= M \left[\left(\frac{X - M(X)}{\sigma_x} \right)^2 \pm 2 \left(\frac{X - M(X)}{\sigma_x} \frac{Y - M(Y)}{\sigma_y} \right) + \left(\frac{Y - M(Y)}{\sigma_y} \right)^2 \right] = 2 \pm 2\rho. \end{aligned}$$

Если $2 \pm 2\rho = 0$, то из того, что у неотрицательной случайной величины математическое ожидание равно нулю, вытекает, что сама величина тождественно равна нулю и

$$Y = M(Y) + \sigma_y / \sigma_x (X - M(X)) \text{ при } \rho = 1$$

и

$$Y = M(Y) - \sigma_y / \sigma_x (X - M(X)) \text{ при } \rho = -1.$$

• 4.3. Двумерные непрерывные случайные величины

Для двумерной непрерывной случайной величины (X, Y) двумерное распределение задается с помощью плотности $p(x, y)$. Определение безусловного распределения компоненты, ее математического ожидания и дисперсии, условной плотности, условного математического ожидания и условной дисперсии аналогичны тем, которые мы привели для дискретных распределений, с той только разницей, что вместо сумм в формулах пишутся интегралы.

В качестве примера рассмотрим двумерную нормальную случайную величину.

Случайная величина (X, Y) распределена по **двумерному нормальному закону**, если совместная плотность величин x и y имеет вид

$$p(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-a_x}{\sigma_x}\right)^2 - 2\rho\frac{x-a_x}{\sigma_x}\frac{y-a_y}{\sigma_y} + \left(\frac{y-a_y}{\sigma_y}\right)^2\right]}.$$

Таким образом, двумерный нормальный закон распределения определяется пятью параметрами: $a_x, a_y, \sigma_x, \sigma_y, \rho$. Параметры a_x, a_y выражают математические ожидания величин X и Y , параметры σ_x, σ_y – их среднеквадратические отклонения, ρ – коэффициент корреляции случайных величин X и Y .

Можно показать, что условные математические ожидания и дисперсии, сосчитанные по этому закону распределения, выражаются с помощью следующих формул:

$$M_y(X) = M(X/Y = y) = a_x + \rho \frac{\sigma_x}{\sigma_y} (y - a_y);$$

$$D_y(X) = D(X/Y = y) = \sigma_x^2 (1 - \rho^2);$$

$$M_x(Y) = M(Y/X = x) = a_y + \rho \frac{\sigma_y}{\sigma_x} (x - a_x);$$

$$D_x(Y) = D(Y/X = x) = \sigma_y^2 (1 - \rho^2).$$

Из этих формул вытекает, что линии регрессии $M_y(X)$ и $M_x(Y)$ нормально распределенных случайных величин представляют собой прямые линии (регрессии Y по X и X по Y линейны) (рис.4.2).

Условные дисперсии постоянны и не зависят от значения второй компоненты. Второе свойство называется **равноизменчивостью** ус-

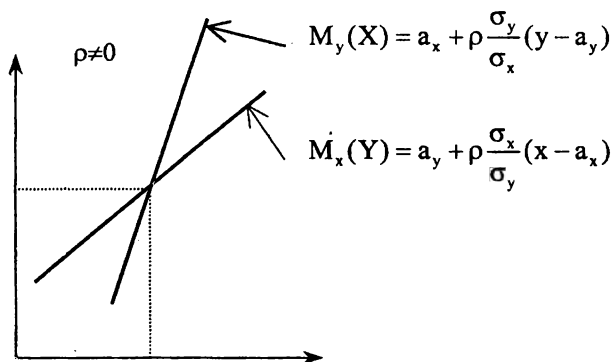


Рис. 4.2

ловных нормальных распределений.

Заметим, что из вида плотности двумерного нормального распределения следует, что если $\rho=0$, то компоненты X и Y независимы. Действительно, в этом случае

$$p(x, y) = \frac{1}{\sigma_x \sqrt{2\pi}} e^{-\frac{(x-a_x)^2}{2\sigma_x^2}} \cdot \frac{1}{\sigma_y \sqrt{2\pi}} e^{-\frac{(y-a_y)^2}{2\sigma_y^2}} = \varphi_1(x) \cdot \varphi_2(y),$$

где $\varphi_1(x)$ и $\varphi_2(y)$ – плотности вероятностей одномерных случайных величин X и Y . Уравнения регрессии обратятся в $M_y(X)=a_x$ и $M_x(Y)=a_y$.

Линии регрессии при $\rho=0$ приобретут вид двух прямых, перпендикулярных координатным осям (рис. 4.3).

С уменьшением ρ ножницы становятся все шире, пока не обратятся в две линии, перпендикулярные осям координат.

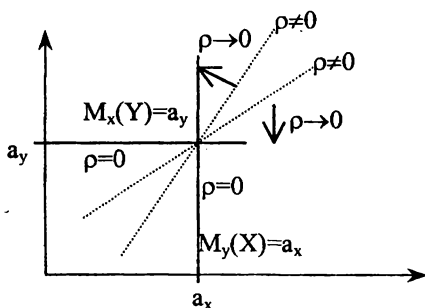


Рис. 4.3

4.4. Описание выборок из двумерных случайных величин

4.4.1. Корреляционные таблицы

При изучении двумерных случайных величин задачей математической статистики, кроме оценки параметров распределения, является определение вида зависимости между компонентами, оценка тесноты связи между ними и построение по выборочным данным оценок для коэффициентов линий регрессии, коэффициента корреляции, нахождение для них доверительных интервалов.

Для двумерных величин выборочные данные представляются в виде таблицы.

Если данных немного, таблицы могут состоять из двух строк:

x_i	1	2	3	4	5	6	7
y_i	5	9	6	10	14	17	15

Если данных много, их группируют. Если одно и то же значение X (группированное) в выборке встретилось n_x раз, значение Y – n_y раз, а пара (x, y) – n_{xy} раз, данные наблюдений (группированные) записывают в виде двумерной (состоит из m строк и k столбцов) таблицы, которую называют **корреляционной** (табл. 4.3).

Здесь Y – количество удобрений на 100 га, X – урожайность, ц/га; на пересечении строки и столбца указано количество хозяйств, в которых при указанном количестве удобрений получен соответствующий урожай. В правом столбце указано, сколько раз встретились конкрет-

Таблица 4.3

$Y \backslash X$	10	12	14	16	18	20	n_y
10	9	4	1				14
30	1	10	9	3			23
50		2	6	14	6		28
70			1	10	18	6	35
n_x	10	16	17	27	24	6	$n=100$

ные значения y , в нижней строке – сколько раз встретились конкретные значения x , в клетке, расположенной в правом нижнем углу, – сумма всех частот (общее число всех наблюдений $n = \sum n_x = \sum n_y$).

Двумерная таблица 4.4 – это таблица эмпирических частот, а если все частоты n_{ij} в ней поделить на число испытаний $n = \sum n_i = \sum n_j$, то получим **двумерную эмпирическую таблицу распределения** (табл. 4.5).

Таблица 4.4

2

Y \ X	x_1	x_2	...	x_k	$\sum_{j=1}^k$
y_1	n_{11}	n_{12}	...	n_{1k}	$n_{1*} = \sum n_{1j}$
y_2	n_{21}	n_{22}	...	n_{2k}	$n_{2*} = \sum n_{2j}$
...	...	...	...	...	...
y_m	n_{m1}	n_{m2}	...	n_{mk}	$n_{m*} = \sum n_{mj}$
$\sum_{i=1}^m$	$n_{*1} = \sum n_{i1}$	$n_{*2} = \sum n_{i2}$	...	$n_{*k} = \sum n_{ik}$	$n = \sum n_{*i} = \sum n_{j*}$

Таблица 4.5

3

Y \ X	x_1	x_2	...	x_k	$\sum_{j=1}^k$
y_1	p'_{11}	p'_{12}	...	p'_{1k}	$p'_{1*} = \sum p'_{1j}$
y_2	p'_{21}	p'_{22}	...	p'_{2k}	$p'_{2*} = \sum p'_{2j}$
...	...	...	...	...	...
y_m	p'_{m1}	p'_{m2}	...	p'_{mk}	$p'_{m*} = \sum p'_{mj}$
$\sum_{i=1}^m$	$p'_{*1} = \sum p'_{i1}$	$p'_{*2} = \sum p'_{i2}$	...	$p'_{*k} = \sum p'_{ik}$	1

4.4.2. Выборочные условные средние

В качестве оценок условных математических ожиданий принимают условные средние, которые находят по данным наблюдений (по выборке).

Условным средним $\bar{y}_x$ называют среднее арифметическое наблюдавшихся значений Y , соответствующих $X=x$. Аналогично условным средним $\bar{x}_y$ называют среднее арифметическое наблюдавшихся значений X , соответствующих $Y=y$.

По эмпирической двумерной таблице распределения выборочные условные средние вычисляются по следующим формулам:

$$\bar{y}_x = M(Y/X = x_j) = 1/p'_{*j} \sum_{i=1}^m y_i p'_{ij};$$

$$\bar{x}_y = M(X/Y = y_i) = 1/p'_{i*} \sum_{j=1}^k x_j p'_{ij}.$$

Если задана корреляционная таблица, то формулы в силу соотношений $p'_{ij} = n_{ij}/n$, $p'_{*j} = \sum p'_{ij} = \sum n_{ij}/n = n_{*j}/n$, $p'_{i*} = \sum p'_{ij} = \sum n_{ij}/n = n_{i*}/n$ принимают вид:

$$\bar{y}_x = M(Y/X = x_j) = 1/n_{*j} \sum_{i=1}^m y_i n_{ij};$$

$$\bar{x}_y = M(X/Y = y_i) = 1/n_{i*} \sum_{j=1}^k x_j n_{ij}.$$

Условное среднее $\bar{y}_x$ так же, как и $M(Y/x)$, является функцией x . Введя обозначение

$$\bar{y}_x = f^*(x),$$

можем сказать, что $f^*(x)$ – **выборочная регрессия Y на X** , а ее график – **выборочная линия регрессии Y на X** .

Аналогично уравнение

$$\bar{x}_y = \varphi^*(y)$$

называют **выборочным уравнением регрессии X на Y** , а его график – **выборочной линией регрессии X на Y** .

Как найти по данным наблюдений вид функций $f'(x)$ и $\phi'(y)$ или их параметры, если вид их известен (например, мы считаем, что регрессия в нашем случае линейная)? Как определить тесноту связи между величинами X и Y и установить, коррелированы ли эти величины? Постараемся ответить на эти вопросы.

5. ЭЛЕМЕНТЫ РЕГРЕССИОННОГО АНАЛИЗА

5.1. Установление формы связи между переменными, прогнозы

Первое представление о наличии и виде связи между признаками дает **поле корреляции**. В случае, когда пар немного, все пары (x, y) наносятся на координатную плоскость. В случае, когда данные сгруппированы, на координатной плоскости рисуют сетку, соответствующую группировке, и в каждом прямоугольнике рисуют столько точек, сколько выборочных пар (x, y) попало в соответствующее гнездо (рис. 5.1, а).

При отсутствии тесных связей имеет место беспорядочное расположение точек (x, y) на графике (рис. 5.1, б, в). Чем сильнее связь между признаками, тем теснее будут группироваться точки вокруг определенной линии, выражающей форму связи.

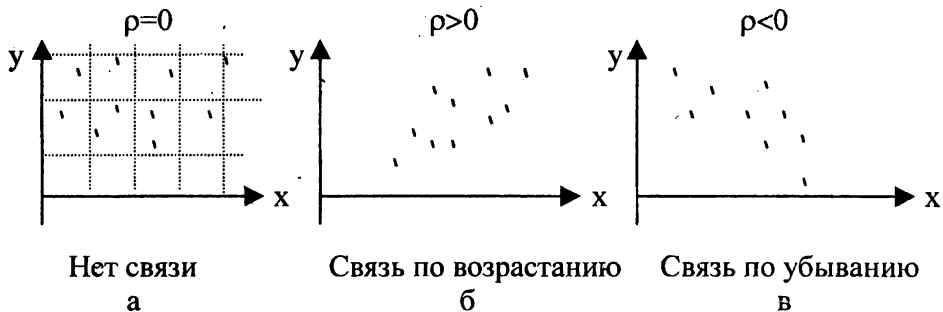


Рис. 5.1

В самом простом случае один из двух признаков, например x , выступает в качестве независимой компоненты (чаще всего это время). Тогда переменная x вообще не случайна, а случайна только переменная y . Например, график данных о прибылях фирмы за предыдущие годы похож на прямую. Надо построить прямую, которая хорошо отражала бы тенденцию изменения прибыли. Такая прямая является эмпирической прямой **регрессии**, в данном случае регрессии y на x .

Когда обе переменные “равноправны”, то строят две прямые регрессии – y на x и x на y (эмпирические прямые регрессии мы в дальнейшем будем называть просто прямыми регрессии). Для построения прямых регрессии применяют метод, который называется методом наименьших квадратов. Для оценки тесноты связи между признаками в числовой форме вычисляют эмпирический коэффициент корреляции – характеристику, выражающую тесноту связи между признаками в числовой форме.

5.2. Метод наименьших квадратов

Метод наименьших квадратов применяется не только для построения прямых. Его можно применять во всех случаях, когда требуется по выборочным данным построить “хорошую” оценку параметров. Суть **метода наименьших квадратов** заключается в том, что оценка параметра определяется из условия минимизации суммы квадратов отклонений выборочных данных от определяемой оценки.

Пример 5.1. Найти оценку для генерального среднего μ методом наименьших квадратов.

Решение. Выпишем минимизируемую функцию (сумма квадратов отклонений от оценки) и условие ее минимизации – равенство нулю ее производной по оценке:

$$u = \sum_{i=1}^n (x_i - \Theta_n)^2, \quad \frac{du}{d\Theta_n} = -2 \sum_{i=1}^n (x_i - \Theta_n) = 0,$$

откуда:

$$\sum_{i=1}^n x_i - n\Theta_n = 0 \quad \text{и} \quad \Theta_n = \frac{\sum_{i=1}^n x_i}{n}.$$

Следовательно, согласно методу наименьших квадратов искомая оценка генеральной средней μ есть выборочное среднее $\bar{X}$.

Огромным преимуществом метода наименьших квадратов является то, что для него не требуется знание законов распределения вероятностей.

Рассмотрим подробно, как применяется метод наименьших квадратов для нахождения параметров прямой в частном случае, когда одна из переменных, например X , не является случайной, а статистическая связь между X и Y является линейной (т.е. связь между ними может

быть приближенно выражена уравнением прямой линии), и задача опыта – найти оценки для параметров прямой, описывающей зависимость Y от X . Совершенно аналогично он применяется в задачах регрессионного анализа для нахождения параметров функций регрессии, вид которых известен и зависимость от параметров является линейной.

5.3. Построение прямой методом наименьших квадратов

Пусть известны результаты опыта, целью которого является исследование зависимости определенной величины от другой (y от x) (например, величины прибыли от объема инвестиций, изменение по месяцам курса доллара и т.д.), т.е. имеются результаты n независимых измерений - опытные точки (x_i, y_i) , где $i = 1, \dots, n$. Мы наносим эти точки на координатную плоскость (рис. 5.2).

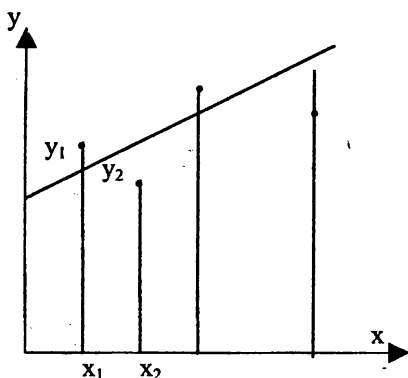


Рис. 5.2

Полученный график дает нам основания предполагать, что статистическая связь между y и x может быть приближенно выражена в виде уравнения:

$$Y(x) = \alpha x + \beta.$$

Среди всех прямых линий $y = ax + b$ на плоскости мы будем искать наиболее близкую к данной системе точек, причем близость будем измерять суммой S квадратов отклонений:

$$S = \sum_{i=1}^n [y_i - (ax_i + b)]^2.$$

То есть из всех прямых мы хотим выбрать ту, для которой сумма S минимальна.

Поскольку минимизируется сумма квадратов разностей экспериментальных и теоретических значений функции (их называют невязками), предложенную процедуру именуют методом наименьших квадратов. А прямую, построенную по этому принципу, для краткости обычно называют **МНК-прямой**.

Эта задача сводится к решению двух уравнений:

$$\begin{cases} \frac{\partial}{\partial a} \sum_{i=1}^n (y_i - ax_i - b)^2 = 0 \\ \frac{\partial}{\partial b} \sum_{i=1}^n (y_i - ax_i - b)^2 = 0 \end{cases} \Rightarrow \begin{cases} -2 \sum_{i=1}^n (y_i - ax_i - b) \cdot x_i = 0; \\ -2 \sum_{i=1}^n (y_i - ax_i - b) = 0. \end{cases}$$

Раскрывая скобки и группируя, в результате получим следующую систему двух линейных уравнений для определения a и b :

$$\begin{cases} \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) \cdot a + \left(\frac{1}{n} \sum_{i=1}^n x_i \right) \cdot b = \frac{1}{n} \sum_{i=1}^n x_i \cdot y_i; \\ \left(\frac{1}{n} \sum_{i=1}^n x_i \right) \cdot a + b = \frac{1}{n} \sum_{i=1}^n y_i. \end{cases}$$

Решая эту систему методом исключения (Гаусса), в итоге получим:

$$a = \frac{\frac{1}{n} \sum x_i y_i - \frac{1}{n} \sum x_i \cdot \frac{1}{n} \sum y_i}{\frac{1}{n} \sum x_i^2 - \left(\frac{1}{n} \sum x_i \right)^2} = \frac{\sum y_i \Delta x_i}{\sum (\Delta x_i)^2} = \frac{\sum y_i \Delta x_i}{\sum x_i \cdot \Delta x_i};$$

$$b = \bar{y} - a \cdot \bar{x}; \quad \bar{x} = \frac{1}{n} \sum x_i; \quad \bar{y} = \frac{1}{n} \sum y_i; \quad \Delta x_i = x_i - \bar{x}.$$

(Во всех суммах знак $\sum$ означает суммирование по всем точкам $\sum_{i=1}^n$).

Уравнение МНК-прямой можно написать и в такой форме:

$$y = a(x - \bar{x}) + \bar{y},$$

откуда видно, что эта прямая проходит через точку $(\bar{x}, \bar{y})$, являющуюся

центром тяжести данной системы точек.

В задачах экономики построенная методом наименьших квадратов линия часто используется для получения представления о динамике процесса и, следовательно, для прогнозов.

Пример 5.2. Динамика производства готовой продукции на фирме отражена в табл. 5.1.

Таблица 5.1

t_i (годы)	1989	1990	1991	1992	1993	1994	1995
y_i (продукция)	18	21	26	22	25	28	30

Поскольку начало отсчета времени мы можем выбрать произвольно, при построении прямой для облегчения счета мы переходим от года к условной единице измерения времени $x = t - 1992$. За ноль мы приняли среднюю точку. Тогда:

$$\bar{x} = 0; \Delta x_i = x_i - \bar{x} = x_i;$$

$$a = \frac{\sum y_i \Delta x_i}{\sum (\Delta x_i)^2} = \frac{49}{28} = 1,75;$$

$$b = \bar{y} - a \cdot \bar{x} = \frac{170}{7} - 1,75 \cdot 0 = 24,286;$$

$$y = 1,75(x - \bar{x}) + 24,286 = 1,75x + 24,286.$$

Таблица 5.2

t_i (годы)	1989	1990	1991	1992	1993	1994	1995	Σ
y_i (продукция)	18	21	26	22	25	28	30	170
$x_i - \Delta x_i$	-3	-2	-1	0	1	2	3	0
$x_i^2 - (\Delta x_i)^2$	9	4	1	0	1	4	9	28
$y_i \cdot \Delta x_i$	-54	-42	-26	0	25	56	90	49
$ax_i + b$	19,04	20,79	22,54	24,29	26,04	27,79	29,54	
$(ax_i + b - y_i)^2$	1,08	0,04	11,83	5,24	1,08	0,04	0,19	19,5

Здесь x измеряется в годах и отсчитывается от 1992 года. Прогнозом на 1996 год будет значение:

$$y = 1,75 \cdot (1996 - 1992) + 24,286 = 7 + 24,286 = 31,284.$$

Такой прогноз будет точечным.

Приведенный пример демонстрирует самую простую схему, позволяющую быстро и просто получить представление о динамике процесса и элементарный прогноз на ближайшее будущее. При решении подобных задач экономического анализа обычно используется значительно больше данных. Перед тем как начать строить методом наименьших квадратов прямую (или другую линию), обычно применяют сглаживание данных методом скользящих средних. О том, как выполняется такое сглаживание, можно прочесть, например, в работе [4]. Для прогнозируемого значения или для значения, соответствующего любой промежуточной точке x , в которой не было произведено измерения, можно построить доверительный интервал. Для того чтобы его построить, опишем вероятностную модель, которой соответствует данная задача построения прямой методом наименьших квадратов. Эта модель имеет следующий вид.

Предположим, что случайная переменная Y зависит от неслучайной независимой переменной X и эта зависимость может быть представлена в виде

$$Y = \alpha X + \beta + \varepsilon,$$

где ε – случайная переменная, которую мы будем называть возмущением.

Это означает, что для измеренных значений y_i можно написать

$$y_i = \beta + \alpha x_i + \varepsilon_i.$$

Предполагаем, что $\varepsilon_i \sim N(0, \sigma)$ и ε_i и ε_j – независимы. Можно сказать, что каждому независимому значению x соответствует распределение вероятностей величины $Y(x)$, для которого условное математическое ожидание $M(Y(x)) = M(Y/X=x) = \alpha x + \beta$.

Тогда построенная нами методом наименьших квадратов прямая $y(x) = \alpha x + \beta$ является оценкой для условного математического ожидания $M(Y(x)) = M(Y/X=x) = \alpha x + \beta$ описанной модели. Оценкой для дисперсии σ^2 является выборочная остаточная дисперсия

$$s^2 = 1/(n-2) \sum_{i=1}^n (ax_i + b - y_i)^2.$$

В знаменателе стоит число степеней свободы $n - 2$, а не n , так как

две степени свободы теряются при определении двух параметров а и b.

Построим доверительный коридор для прямой МНК. Для этого напишем ее уравнение в виде

$$y(x) = a(x - \bar{x}) + \bar{y}.$$

Дисперсия величины $y(x)$ в точке x вычисляется в соответствии со свойствами дисперсии по следующей формуле:

$$\sigma_{y(x)}^2 = \sigma_{\bar{y}}^2 + \sigma_a^2(x - \bar{x})^2.$$

При этом

$$\sigma_{\bar{y}}^2 = \frac{\sigma^2}{n}; \quad \sigma_a^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2}{\left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \text{так как } a = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Отсюда оценка дисперсии значения $y(x)$ в точке x :

$$s_{y(x)}^2 = s^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right).$$

Исходя из того, что статистика

$$t = \frac{y(x) - (ax + b)}{s_{y(x)}}$$

имеет распределение Стьюдента с $n - 2$ степенями свободы, доверительный интервал для среднего значения $y(x)$ в точке x имеет вид

$$ax + b - t_{1-\alpha; n-2} \cdot s_{y(x)} < y(x) < ax + b + t_{1-\alpha; n-2} \cdot s_{y(x)}.$$

Построенная доверительная область для прямой МНК определяет доверительный интервал для математического ожидания, а не для самого случайного значения $Y(x)$. При определении доверительного интервала для индивидуальных значений $Y(x)$ необходимо учесть еще один источник вариации – рассеяние вокруг математического ожидания (величину ϵ). Это означает, что для значений $Y(x)$ в оценку суммарной дисперсии следует включить величину s^2 :

$$s_{Y(x)}^2 = s^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right],$$

и соответствующий доверительный интервал для прогнозов индивидуальных значений $Y(x)$ будет определяться по формуле

$$ax + b - t_{1-\alpha; n-2} \cdot s_{Y(x)} < Y(x) < ax + b + t_{1-\alpha; n-2} \cdot s_{Y(x)}.$$

Пример 5.3. Найдем 95%-ные доверительные интервалы для прогнозируемого среднего и индивидуального значения объема готовой продукции в 1996 году (см. пример 5.2):

$$s^2 = 19,5 / (7 - 2) = 3,9;$$

$$s_{y(x)}^2 = s^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) = 3,9 \left(\frac{1}{7} + \frac{4^2}{28} \right) = 2,769; \quad s_{y(x)} = 1,664;$$

$$s_{Y(x)}^2 = s^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) = 3,9 \left(1 + \frac{1}{7} + \frac{4^2}{28} \right) = 6,67; \quad s_{Y(x)} = 2,58;$$

$$t_{0,95;5} = 2,57;$$

$$31,284 - 2,57 \cdot 1,664 < M(Y(1996)) < 31,284 + 2,57 \cdot 1,664;$$

$$31,284 - 4,276 < M(Y(1996)) < 31,284 + 4,276;$$

$$27,008 < M(Y(1996)) < 35,56;$$

$$31,284 - 2,57 \cdot 2,58 < Y(1996) < 31,284 + 2,57 \cdot 2,58;$$

$$31,284 - 6,64 < Y(1996) < 31,284 + 6,64;$$

$$24,64 < Y(1996) < 37,92.$$

5.4. Проверка значимости уравнения регрессии

Проверить **значимость уравнения регрессии** – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным.

Проверка проводится на основе дисперсионного анализа.

Согласно основной идее дисперсионного анализа

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (ax_i + b - \bar{y})^2 + \sum_{i=1}^n [y_i - (ax_i + b)]^2 = Q_R + Q_{\text{ост}}$$

или

$$Q = Q_R + Q_{\text{ост}},$$

где Q – общая сумма квадратов отклонений зависимой переменной от общей средней, а Q_R и $Q_{\text{ост}}$ – соответственно сумма квадратов отклонений, обусловленная регрессией, и остаточная сумма квадратов, характеризующая отклонение от среднего в точке x_i , т.е. влияние неучтенных факторов.

Схема дисперсионного анализа описывается табл. 5.3.

Таблица 5.3

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Средние квадраты
Регрессия	Q_R	$m - 1$	$s^2_R = Q_R / (m - 1)$
Остаточная	$Q_{\text{ост}}$	$n - m$	$s^2 = Q_{\text{ост}} / (n - m)$
Общая	Q	$n - 1$	

Уравнение регрессии значимо на уровне α , если фактически наблюдаемое значение статистики

$$F = \frac{Q_R (n - m)}{Q_{\text{ост}} (m - 1)} = \frac{Q_R (n - 2)}{Q_{\text{ост}} (2 - 1)} > F_{\alpha; 1; n-2},$$

где $F_{\alpha; 1; n-2}$ – табличное значение F-критерия Фишера-Снедекора, определенное на уровне значимости α ; n – число наблюдений. Число m оцениваемых по выборке параметров в нашем случае равно двум (a и b). Значение F показывает, в какой мере регрессия лучше оценивает значение зависимой переменной по сравнению с ее средней (именно

поэтому в данном случае проверяемая гипотеза проходит при попадании статистики F в критическую область).

Пример 5.4. Оценим значимость уравнения регрессии для примера об объеме готовой продукции:

$$a = 1,75; b = \bar{y} - a \cdot \bar{x} = \bar{y}, \text{ так как } \bar{x} = 0;$$

$$Q_R = \sum_{i=1}^n (ax_i + b - \bar{y})^2 = a^2 \sum_{i=1}^n x_i^2 = (1,75)^2 \cdot 28 = 85,75; Q_{\text{ост}} = 19,5;$$

$$F = 85,75 \cdot (7-2) / 19,5 = 21,99;$$

$$F_{0,05;1;5} = 6,61.$$

Поскольку $F > F_{0,05;1;5}$ ($21,99 > 6,61$), то уравнение регрессии значимо – результаты опыта не противоречат модели $Y = \alpha x + \beta + \varepsilon$ и в качестве оценки можно принять

$$Y = 1,75x + 24,286 + \varepsilon = 1,75(t-1992) + 24,286 + \varepsilon.$$

5.5. Нелинейная регрессия

В случаях, когда зависимость между факторами нельзя выразить линейными функциями, используют **нелинейную** регрессию. Выбор вида уравнения регрессии производится на основе опыта предыдущих исследований, теоретических соображений, а также на основе наблюдения расположения точек корреляционного поля. Наиболее часто встречаются следующие виды уравнений нелинейной регрессии: полиномиальное – $y = b_0 + b_1x + \dots + b_kx^k$; гиперболическое – $y = b_0 + b_1/x$; степенное – $y = b_0 \cdot x^{b_1}$.

Пример 5.5. Данные по исследованию зависимости урожайности зерновых y_i (ц/га) от количества осадков x_i (см) представлены в таблице:

x_i	25	27	30	35	36	38	39	41	42	45	46	47	50	52	53
y_i	23	24	27	27	32	31	33	35	34	32	29	28	25	24	25

По таблице построен график (рис. 5.3).

Если исследуемый показатель y состоит из двух частей – постоянной (не зависящей от x) и переменной, уменьшающейся с ростом x , то зависимость y от x можно представить в виде гиперболы $y = b_0 + b_1/x$. Зависимость вида $y = \frac{a}{x} + b$ – линейная относительно параметров a и b .

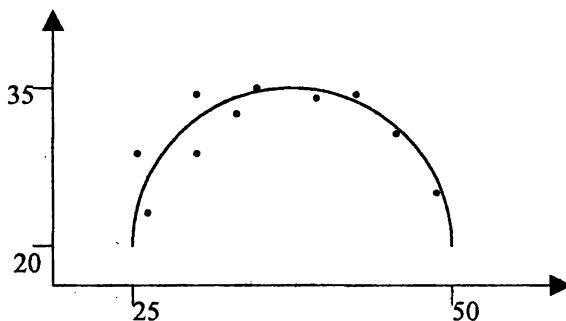


Рис. 5.3

В этом случае задача легко может быть сведена к предыдущей заменой переменной : $u = \frac{1}{x}$.

Если показатель отражает процесс, который под влиянием фактора x происходит с постоянным ускорением или замедлением, то применяются полиномы. При исследовании степенного уравнения регрессии следует иметь в виду, что оно нелинейно относительно параметров b_i , однако путем логарифмирования может быть преобразовано в линейное:

$$\ln y = \ln b_0 + b_1 \ln x.$$

Для определения параметров b_i используется метод наименьших квадратов. Все рассуждения сохраняются.

6. КОРРЕЛЯЦИОННАЯ МОДЕЛЬ

6.1. Две регрессионные кривые

Когда обе переменные x и y «равноправны», т.е. в случае, когда имеет место **корреляционная модель**, строят две регрессионные прямые.

Пример 6.1. Изучается зависимость между суточной выработкой продукции Y и величиной основных производственных фондов (ОПФ) X для 50-ти однотипных предприятий. Данные обследования представлены в корреляционной табл. 6.1.

Таблица 6.1

ОПФ, млн. руб.	Середины интерва- лов	Суточная выработка продукции Y, т						$\Sigma = n_{i*}$	$\bar{y}_i$
		7-11	11-15	15-19	19-23	23-27			
	$x_i \backslash y_j$	9	13	17	21	25			
20-25	22,5	2	1	-	-	-	3	10,3	
25-30	27,5	3	6	4	-	-	13	13,3	
30-35	32,5	-	3	11	7	-	21	17,8	
35-40	37,5	-	1	2	6	2	11	20,3	
40-45	42,5	-	-	-	1	1	2	23,0	
$\Sigma = n_{*j}$		5	11	17	14	3	50		
$\bar{x}_j$		25,5	29,3	31,9	35,4	39,2			

Для каждой строки (т.е. для каждого значения x_i) вычислим групповые средние (условные математические ожидания):

$$\bar{y}_i = 1/n_{i*} \sum_{j=1}^m y_j n_{ij}, \text{ где } m - \text{число интервалов группировки для } Y.$$

Вычисленные групповые средние $\bar{y}_i$ мы поместили в последнем столбце корреляционной таблицы 6.1. Точки $(x_i, \bar{y}_i)$ обозначим на координатной плоскости. Соединив их прямолинейными отрезками, получим эмпирическую линию регрессии Y на X (рис. 6.1). Она, вообще говоря, не является прямой.

С помощью метода наименьших квадратов мы можем построить для точек $(x_i, \bar{y}_i)$ прямую, которая и будет оценкой генеральной линии регрессии Y на X .

Будем искать коэффициенты a и b такие, что

$$S = \sum_{i=1}^k (ax_i + b - \bar{y}_i)^2 n_{i*} = \min \quad (k - \text{число интервалов группировки для } X).$$

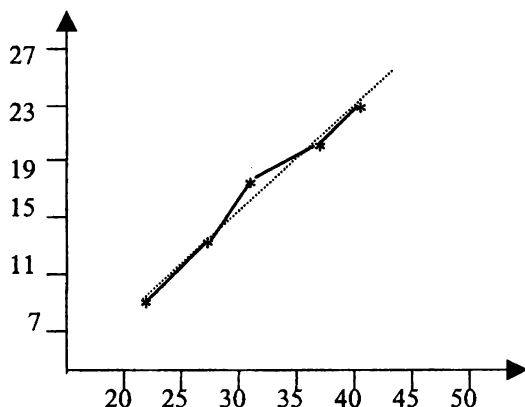


Рис. 6.1

Такую задачу мы уже решали, когда проводили МНК-прямую по данным, заданным таблицей, состоящей из двух строк. Только теперь во всех формулах появятся веса, задающие число одинаковых слагаемых. Введем естественные обозначения:

$$\bar{x} = 1/n \sum_{i=1}^k x_i n_{i*}; \quad \bar{y} = 1/n \sum_{j=1}^m y_j n_{*j}; \quad \overline{xy} = 1/n \sum_{i=1}^k \sum_{j=1}^m x_i y_j n_{ij}; \quad \overline{x^2} = 1/n \sum_{i=1}^k x_i^2 n_{i*};$$

Уравнение прямой регрессии Y на X имеет вид

$$y = a_{yx}(x - \bar{x}) + \bar{y},$$

где $a_{yx} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\mu}{s_x^2};$

$$\mu = \overline{xy} - \bar{x} \cdot \bar{y} = \frac{\sum_{i=1}^k \sum_{j=1}^m x_i y_j n_{ij}}{n} - \bar{x} \cdot \bar{y} \quad - \text{выборочная ковариация};$$

$$s_x^2 = \frac{\sum_{i=1}^k x_i^2 n_{i*}}{n} - (\bar{x})^2 \quad - \text{выборочная дисперсия переменной } X.$$

В силу равноправия величин X и Y и симметрии формул относительно X и Y прямая регрессии x на y строится по формулам, в которых x и y поменяли местами.

Для того чтобы понять, что дает вид этих двух прямых, представим их уравнения в следующей форме:

$$y_x - \bar{y} = a_{yx}(x - \bar{x});$$

$$x_y - \bar{x} = a_{xy}(y - \bar{y}).$$

Коэффициенты a_{yx} и a_{xy} называются **выборочными коэффициентами регрессии** (у на x и x на y соответственно). Коэффициент регрессии показывает, на сколько единиц в среднем изменяется одна величина относительно другой, когда та меняется на единицу.

Например, $a_{yx} = 2$. Когда x меняется на 1, y в среднем меняется на 2.

Сравним формулы, по которым вычисляются коэффициенты регрессии:

$$a_{yx} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{s_x^2} = \frac{\mu}{s_x^2};$$

$$a_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{s_y^2} = \frac{\mu}{s_y^2},$$

где s_x^2 и s_y^2 – выборочные дисперсии переменных X и Y, а через μ мы обозначили выборочный корреляционный момент или выборочную ковариацию:

$$\mu = \overline{xy} - \bar{x} \cdot \bar{y}.$$

Вид этих уравнений показывает, что коэффициенты регрессии a_{yx} и a_{xy} имеют одинаковые знаки, определяемые знаком μ . Из уравнений регрессии следует, что коэффициенты a_{yx} и $1/a_{xy}$ определяют угловые коэффициенты (тангенсы углов наклона) к оси Oх соответствующих линий регрессии, пересекающихся в точке $(\bar{x}, \bar{y})$ (рис. 6.2).

Угол между прямыми регрессии является важным показателем зависимости компонент. Мы уже говорили о числовой характеристике, выражающей тесноту связи между признаками в числовой форме. Такой характеристикой является коэффициент корреляции. Покажем, как связаны угол между прямыми регрессии и коэффициент корреляции.

6.2. Эмпирический коэффициент корреляции

Когда наблюдения проводятся над системой (X,Y) двух случайных величин, то для оценки тесноты связи между ними по результатам вы-

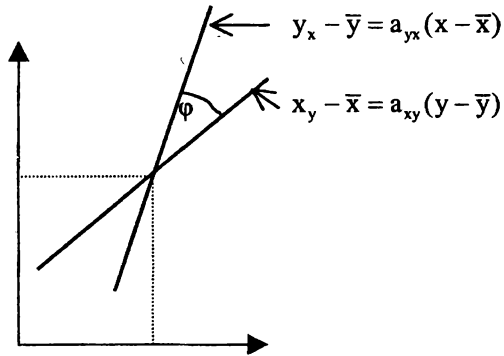


Рис. 6.2

борки строится статистика, называемая **эмпирическим коэффициентом корреляции**. Для таблицы данных, состоящей из двух строк, имеющей вид:

x_i	x_1	x_2	...	x_n
y_i	y_1	y_2	...	y_n

коэффициент корреляции вычисляется по формуле:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

или, что то же:

$$r_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / s_x \cdot s_y$$

Эту формулу легко преобразовать к виду

$$r_{xy} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i y_i) - \bar{x} \cdot \bar{y}}{\sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{y}^2}} = \frac{\mu}{s_x \cdot s_y}$$

По виду формулы легко заметить, что в нее входят нормированные значения обеих компонент – из каждого значения вычитается среднее и разность делится на среднеквадратическое отклонение. Эта операция продлевается для того, чтобы избавиться от влияния сдвига и выбора масштаба, в котором измеряется компонента. Так что нормировка переводит каждый ряд значений в шкалу, у которой нулем считается среднеарифметическое исходных цифр, а за единицу принята величина среднеквадратического отклонения.

Если выборочные данные заданы в виде двумерной табл. 6.2,

Таблица 6.2

$\begin{matrix} Y \\ \backslash \\ X \end{matrix}$	x_1	x_2	$\dots$	x_k	$\sum_{j=1}^k$
y_1	n_{11}	n_{12}	$\dots$	n_{1k}	$n_{1*} = \sum n_{1j}$
y_2	n_{21}	n_{22}	$\dots$	n_{2k}	$n_{2*} = \sum n_{2j}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
y_m	n_{m1}	n_{m2}	$\dots$	n_{mk}	$n_{m*} = \sum n_{mj}$
$\sum_{i=1}^m$	$n_{*1} = \sum n_{i1}$	$n_{*2} = \sum n_{i2}$	$\dots$	$n_{*k} = \sum n_{ik}$	$n = \sum n_{*i} = \sum n_{j*}$

то при вычислении коэффициента корреляции, задающегося формулой

$$r_{xy} = \frac{\mu}{s_x \cdot s_y},$$

величины μ , s_x и s_y надо вычислять по формулам, в которых учитывается кратность значений:

$$\mu = \overline{xy} - \bar{x} \cdot \bar{y} = \frac{\sum_{i=1}^k \sum_{j=1}^m x_i y_j n_{ij}}{n} - \bar{x} \cdot \bar{y} \quad - \text{выборочная ковариация};$$

$$s_x^2 = \frac{\sum_{i=1}^k x_i^2 n_{*i}}{n} - (\bar{x})^2 - \text{выборочная дисперсия переменной } X;$$

$$s_y^2 = \frac{\sum_{j=1}^m y_j^2 n_{*j}}{n} - (\bar{y})^2 - \text{выборочная дисперсия переменной } Y.$$

Свойства эмпирического коэффициента корреляции. Значения выборочного коэффициента корреляции r_{xy} заключены между -1 и $+1$. Если точки (x_i, y_i) лежат строго на прямой, то есть имеет место строгая линейная зависимость между выборочными значениями X и Y , то $r_{xy} = \pm 1$. В этом можно убедиться, подставив в формулу для r_{xy} значения $y_i = ax_i + b$.

Пример 6.2. Ниже приведены идеальные данные о росте и весе людей среднего возраста, сохранивших “спортивный вес”, т.е. вес, определяемый по формуле: вес = рост – 102:

Рост x_i	178	166	172	168	176
Вес y_i	76	64	70	66	74

Вычислите по формуле коэффициент корреляции и убедитесь, что он равен 1.

Случай, когда эмпирическая корреляция оказывается близкой по модулю к 1, указывает на то, что компоненты связаны друг с другом и связь близка к линейной зависимости.

Пример 6.3. В табл. 6.3 приведены данные о производительности труда (Y – в т/ч) и уровне механизации работ (X – в процентах) для 14 предприятий. Требуется установить зависимость между производительностью труда и уровнем механизации работ.

Таблица 6.3

N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Σ
X, %	32	30	36	40	41	47	56	54	60	55	61	67	69	76	724
Y, т/ч	20	24	28	30	31	33	34	37	38	40	41	43	45	48	492
$ax+b$	24,4	23,3	26,6	28,8	29,3	32,6	37,5	36,4	39,6	36,9	40,2	43,5	44,5	48,3	
Δy	4,4	-0,7	-1,4	-1,2	-1,7	-0,4	3,5	-0,6	1,6	-3,1	-0,8	0,5	-0,5	0,3	

$$\bar{x} = \frac{1}{14}(32 + 30 + \dots + 76) = 51,71;$$

$$\bar{y} = \frac{1}{14}(20 + 24 + \dots + 48) = 35,14;$$

$$\frac{1}{n} \sum_{i=1}^n x_i y_i = \frac{1}{14}(32 \cdot 20 + 30 \cdot 24 + \dots + 76 \cdot 48) = 1921,93;$$

$$\frac{1}{n} \sum_{i=1}^n x_i^2 = \frac{1}{14}(32 \cdot 32 + 30 \cdot 30 + \dots + 76 \cdot 76) = 2866,71;$$

$$\frac{1}{n} \sum_{i=1}^n y_i^2 = \frac{1}{14}(20 \cdot 20 + 24 \cdot 24 + \dots + 48 \cdot 48) = 1295,57;$$

$$\begin{aligned} r &= \frac{1921,93 - 51,71 \cdot 35,14}{\sqrt{(2866,71 - 51,71^2)(1295,57 - 35,14^2)}} = \\ &= \frac{104,14}{\sqrt{192,79 \cdot 60,75}} = \frac{104,14}{108,22} = 0,96. \end{aligned}$$

Величина r близка к единице. Следовательно, между величинами существует линейная корреляционная зависимость. Построим по этим точкам прямую методом наименьших квадратов:

$$a = \frac{\frac{1}{n} \sum x_i y_i - \frac{1}{n} \sum x_i \cdot \frac{1}{n} \sum y_i}{\frac{1}{n} \sum x_i^2 - \left(\frac{1}{n} \sum x_i \right)^2} = \frac{104,14}{192,8} = 0,54;$$

$$b = \bar{y} - a \cdot \bar{x} = 7,04.$$

Следовательно, уравнение прямой имеет вид:

$$y = 0,54x + 7,04$$

и можно сделать вывод, что производительность труда в среднем возрастает на 0,54 т/ч, если коэффициент механизации работ увеличится на один процент. Вычислив значения $ax+b$, можно посмотреть, на каких предприятиях отклонения положительные, на каких – отрицатель-

ные, на каких – самые большие.

В предположении, что генеральный коэффициент корреляции $\rho=0$, статистика

$$t = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2}$$

имеет распределение Стьюдента с $(n - 2)$ степенями свободы (два средних заменены на эмпирические значения). Это обстоятельство можно использовать для проверки гипотезы о равенстве ρ нулю. Если выполняется $|t| > t_{1-\alpha, n-2}$, где $t_{1-\alpha, n-2}$ – табличное значение критерия Стьюдента, определенное на уровне значимости α для $(n - 2)$ степеней свободы, то выборочный коэффициент r значимо (существенно) отличается от нуля. Если гипотеза не проходит, то для построения доверительного интервала для истинного коэффициента корреляции ρ , для которого эмпирический коэффициент является точечной оценкой, при $n > 50$ для интервальной оценки генерального коэффициента ρ можно воспользоваться формулой

$$r - k_{1-\alpha} \frac{\sqrt{1-r^2}}{\sqrt{n-2}} \leq \rho \leq r + k_{1-\alpha} \frac{\sqrt{1-r^2}}{\sqrt{n-2}},$$

где $k_{1-\alpha}$ ищется по таблицам нормального распределения (таблица 8 приложения 2)

Например, для уровня значимости $\alpha=0,0025$ интервальная оценка имеет вид

$$r - 3 \frac{\sqrt{1-r^2}}{\sqrt{n-2}} \leq \rho \leq r + 3 \frac{\sqrt{1-r^2}}{\sqrt{n-2}}.$$

При небольших n формула более сложная (ее можно найти, например, в работе [2]).

Близость коэффициента корреляции к нулю не говорит прямо о независимости величин, но является одним из ее признаков. При близости эмпирического коэффициента к нулю можно с помощью доверительного интервала, построенного по распределению Стьюдента, проверить гипотезу о равенстве нулю истинного коэффициента корреляции.

Выборочный коэффициент корреляции r является оценкой генерального коэффициента корреляции ρ , тем более точной, чем больше объем выборки n . Для двумерного нормального распределения равенство нулю коэффициента корреляции означает независимость компо-

нент X и Y . Для такого распространенного случая, как совместное нормальное распределение компонент X и Y , равенство нулю коэффициента корреляции очень информативно.

Посмотрим, что нам дает значение коэффициента корреляции при рассмотрении линий регрессии. Для этого с помощью простых тождественных преобразований выпишем уравнения регрессии через коэффициент корреляции:

$$y_x - \bar{y} = a_{yx}(x - \bar{x}) = r \frac{s_y}{s_x}(x - \bar{x});$$

$$x_y - \bar{x} = a_{xy}(y - \bar{y}) = r \frac{s_x}{s_y}(y - \bar{y}).$$

Мы учли, что:

$$r_{xy} = \frac{\mu}{s_x \cdot s_y};$$

$$a_{yx} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{s_x^2} = \frac{\mu}{s_x^2} = r \frac{s_y}{s_x};$$

$$a_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{s_y^2} = \frac{\mu}{s_y^2} = r \frac{s_x}{s_y}.$$

Легко видеть, что

$$r = a_{yx} \frac{s_x}{s_y} = a_{xy} \frac{s_y}{s_x} = \pm \sqrt{a_{yx} a_{xy}},$$

т.е. коэффициент корреляции r переменных X и переменных Y есть средняя геометрическая коэффициентов регрессии, имеющая их знак.

Пример 6.4. Результаты измерения роста и веса 50-ти человек сгруппированы и занесены в табл. 6.4. По ней изучается зависимость между весом и ростом человека.

$$\bar{x} = 1/50(160 \cdot 12 + 170 \cdot 23 + 180 \cdot 15) = 8530/50 = 170,6;$$

$$\bar{y} = 1/50(60 \cdot 4 + 70 \cdot 17 + 80 \cdot 19 + 90 \cdot 10) = 3850/50 = 77;$$

Таблица 6.4

X(см)\Y(кг)	60	70	80	90	
160	2	5	4	1	12
170	2	8	9	4	23
180	0	4	6	5	15
	4	17	19	10	50

$$s_x^2 = 1/50(160 \cdot 160 \cdot 12 + 170 \cdot 170 \cdot 23 + 180 \cdot 180 \cdot 15) - 170,6 \cdot 170,6 =$$

$$= 1457900/50 - 29104,36 = 53,64; \quad s_x = 7,324;$$

$$s_y^2 = 1/50(60 \cdot 60 \cdot 4 + 70 \cdot 70 \cdot 17 + 80 \cdot 80 \cdot 19 + 90 \cdot 90 \cdot 10) - 77 \cdot 77 =$$

$$= 300300/50 - 5929 = 6006 - 5929 = 77; \quad s_y = 8,775;$$

$$\mu = 1/50(160 \cdot 60 \cdot 2 + 160 \cdot 70 \cdot 5 + 160 \cdot 80 \cdot 4 + 160 \cdot 90 \cdot 1 + 170 \cdot 60 \cdot 2 + 170 \cdot 70 \cdot 8 +$$

$$+ 170 \cdot 80 \cdot 9 + 170 \cdot 90 \cdot 4 + 180 \cdot 60 \cdot 0 + 180 \cdot 70 \cdot 4 + 180 \cdot 80 \cdot 6 + 180 \cdot 90 \cdot 5) - 170,6 \cdot 77 =$$

$$= 1/50(160 \cdot 880 + 170 \cdot 1760 + 180 \cdot 1210) - 13136,2 = 657800/50 =$$

$$= 13156 - 13136,2 = 19,8;$$

$$a_{yx} = m/s_x^2 = 19,8/53,64 = 0,369;$$

$$y = 77 + 0,369(x - 170,6) = 0,369x + 14,05.$$

Вычисляем средний вес для роста 180 см:

$$y(180) = 0,369 \cdot 180 + 14,05 = 66,42 + 14,05 = 80,5.$$

Проводим прямую регрессии веса по росту через точки (170,6; 77) и (180; 80,5):

$$a_{xy} = \mu/s_y^2 = 19,8/77 = 0,257; \quad 1/a_{xy} = 1/0,257 = 3,891;$$

$$x = 170,6 + 0,257(y - 77) = 0,257y + 150,81.$$

Вычисляем средний рост для веса 90 кг:

$$x(90) = 0,257 \cdot 90 + 150,81 = 173,94.$$

Проводим прямую регрессии роста по весу через точки (170,6; 77) и (174; 90) (рис. 6.3).

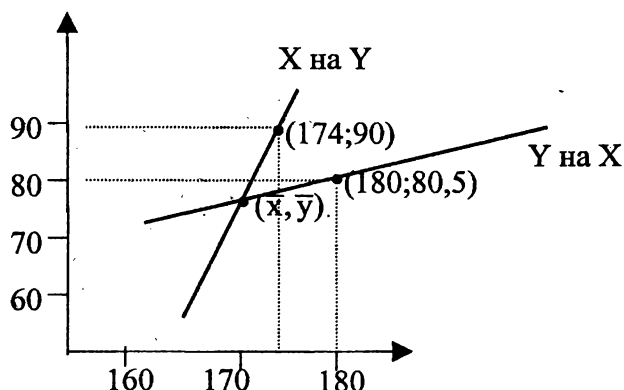


Рис. 6.3

$$r = \mu/s_x s_y = 19,8/8,775 \cdot 7,324 = 0,308.$$

Найдем $\operatorname{tg} \varphi$, где φ – угол между прямыми регрессии. Их угловые коэффициенты соответственно равны $k_1 = a_{yx}$, $k_2 = 1/a_{xy}$. В частности, в нашем примере $k_1 = 0,369$, $k_2 = 1/0,257 = 3,891$.

Воспользовавшись формулой аналитической геометрии, с помощью элементарных выкладок получаем

$$\operatorname{tg} \varphi = \frac{k_2 - k_1}{1 + k_1 k_2} = \frac{1 - a_{yx} a_{xy}}{a_{xy} + a_{yx}} = \frac{1 - r^2}{r} \cdot \frac{s_x s_y}{s_x^2 + s_y^2}.$$

Из полученной формулы видно, что чем ближе $|r|$ к 1, тем меньше угол φ – угол между прямыми регрессии (уже образуемые ими ножницы). При $|r| = 1$ $\operatorname{tg} \varphi = 0$, следовательно, $\varphi = 0$ и линии регрессии сливаются (рис. 6.4, 6.5).

При $r=0$, $a_{yx} = a_{xy} = 0$ линии регрессии обращаются в перпендикулярные прямые: $y_x = \bar{y}$; $x_y = \bar{x}$ (рис. 6.6).

Интересно сравнить формулы прямых регрессии для двумерного нормального распределения с формулами эмпирических линий регрессии. Такое сравнение показывает, что прямые регрессии, построенные методом наименьших квадратов, совпадают с теми, которые мы бы построили, воспользовавшись эмпирическими значениями моментов, то есть по методу моментов.

При r , близком к нулю, **линейная** корреляционная связь практически отсутствует, но это не означает отсутствие зависимости другого

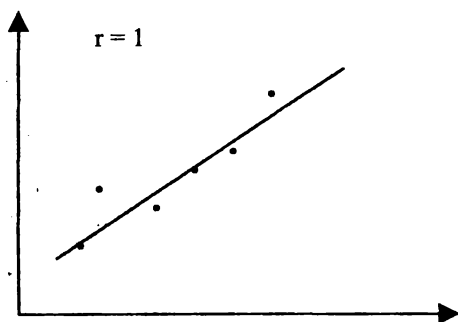


Рис. 6.4

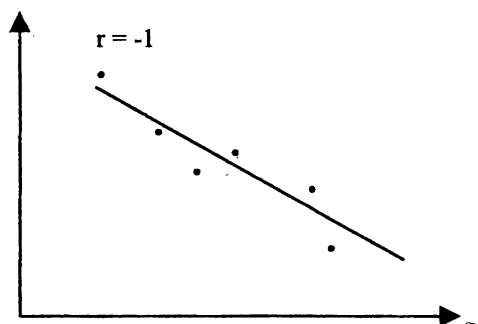


Рис. 6.5

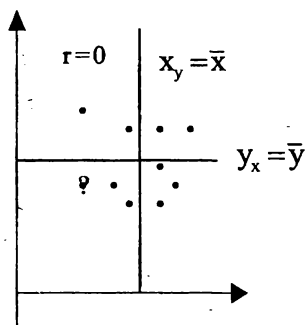


Рис. 6.6

типа. Например, условные математические ожидания с изменением x не меняются, а меняется рассеяние точек поля относительно линии регрессии или регрессия является квадратичной, то есть условные математические ожидания лежат на параболе.

6.3. Множественная корреляция

Если исследуется связь между несколькими признаками (больше двух), то корреляцию называют **множественной**.

В простейшем случае число признаков равно трем. Если связь между ними приблизительно линейная, то можно записать

$$z \sim ax + by + c.$$

В этом случае возникают задачи:

– найти по данным наблюдений выборочное уравнение связи вида

$$z = Ax + By + C;$$

– оценить тесноту связи между признаками.

Первая задача решается методом наименьших квадратов, вторая вычислением соответствующих коэффициентов корреляции. Новым является только понятие выборочного совокупного коэффициента корреляции, оценивающего тесноту связи между Z и обоими признаками X и Y . Смысл всех формул и рассуждений сохраняются. Соответствующие формулы можно найти, например, в работах [1,2].

ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

1. Составьте логическую схему базы знаний по теме юниты.

2. Задачи для самостоятельного решения.

5 ВАРИАНТОВ

Для заданий 1, 2, 3 – номер варианта входит в условие задачи

1. Из таблицы 11 чисел выборки из равномерного распределения на отрезке $[0, 100]$ (см. приложение 2) возьмите подряд 50 чисел, начиная с номера $20(N-1)$, где N – ваш номер варианта (дойдя до конца таблицы, перейдите в ее начало). Возьмите в качестве интервалов группировки интервалы $(0, 20)$, $(20, 40)$... $(80, 100)$ и напишите таблицу эмпирического распределения для этих интервалов. По этой же таблице сосчитайте несмещенную оценку для генеральной дисперсии σ^2 . Постройте для генеральной дисперсии 95%-ный доверительный интервал. Попало ли оцениваемое значение в доверительный интервал? Выпишите теоретические значения для вероятностей попадания в интервалы $(0, 20)$, $(20, 40)$... $(80, 100)$ и сравните их с эмпирическими, т.е. с помощью критерия χ^2 проверьте гипотезу о равномерном распределении $R(0, 100)$ генеральной совокупности (уровень значимости 0,05).

2. Из таблицы 12 чисел выборки из нормального распределения $N(0, 1)$ (см. приложение 2) возьмите подряд 50 чисел, начиная с номера $30(N-1)$, где N – ваш номер варианта (дойдя до конца таблицы, перейдите в ее начало). Возьмите в качестве интервалов группировки интервалы: $(-3, -1)$, $(-1, 0)$, $(0, 1)$, $(1, 3)$ и напишите таблицу эмпирического распределения для этих интервалов. По этой же таблице сосчитайте несмещенную оценку для генеральной дисперсии σ^2 . Постройте для

генеральной дисперсии 95%-ный доверительный интервал. Попало ли оцениваемое значение в доверительный интервал? Выпишите теоретические значения для вероятностей попадания в интервалы $(-3, -1)$, $(-1, 0)$, $(0, 1)$, $(1, 3)$ (для этого вспомните правило трех σ , правило двух σ , правило одного σ или воспользуйтесь таблицей 7) и сравните их с эмпирическими, т.е. с помощью критерия χ^2 проверьте гипотезу о нормальном распределении $N(0, 1)$ генеральной совокупности (уровень значимости 0,05).

3. Постройте таблицу распределения такую же, как в задании 1, взяв 50 чисел из таблицы 11, не вошедших в ваши 50 чисел задания 1, и проверьте гипотезу об однородности этих выборок.

Для заданий 4, 5, 6 даны 5 вариантов

Вариант 1

4. Произведено по пять испытаний на каждом из четырех уровней фактора F . Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве средних для разных уровней фактора. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Результаты измерений приведены в таблице:

Номер испытания	Уровни фактора			
i	F_1	F_2	F_3	F_4
1	36	56	52	39
2	47	51	57	57
3	50	64	59	63
4	58	66	58	61
5	67	66	79	65
$\Sigma/5$	51,6	62,6	61,0	57,0

Указание. Принять $y_{ij} = x_{ij} - 58$.

5. В таблице приведены выборочные данные о количестве поглощенного крысами кислорода (Y) в зависимости от температуры внешней среды (X).

x_i	0	5	10	15	20	25	30
y_i	3,8	3,4	2,6	2,0	1,7	1,4	1,3

Найти уравнение регрессии Y на X . Построить его график.

6. В таблице приведены данные о количестве внесенных удобрений X (в ц /га) и урожайности Y (в ц/га) на 100 га пахотной земли.

$X \backslash Y$	10	12	14	16	18	20
10	9	4	1			
30	1	10	9	3		
50		2	6	14	6	
70			1	10	18	6

Составить уравнения прямых регрессий Y на X и X на Y , построить их графики. Какой будет средний урожай, если внести по 80 центнеров удобрений на гектар? Сколько надо вносить удобрений, чтобы получать урожай 15 центнеров с гектара?

Вычислить коэффициент корреляции.

Вариант 2

4. Произведено по шесть испытаний на каждом из пяти уровней фактора F . Методом дисперсионного анализа при уровне значимости 0,1 проверить нулевую гипотезу о равенстве средних для разных уровней фактора. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Результаты измерений приведены в таблице:

Номер испытания	Уровни фактора				
i	F_1	F_2	F_3	F_4	F_5
1	12	6	9	17	10
2	11	11	6	14	14
3	10	8	8	13	11
4	13	7	10	15	9
5	11	7	10	12	12
6	8	9	11	13	15
Σ	65	48	54	84	71

Указание. Принять $y_{ij} = x_{ij} - 10$.

5. Найти уравнение прямой регрессии Y на X и построить его график по данным пяти наблюдений:

x_i	1,00	1,50	3,00	4,50	5,00
y_i	1,25	1,40	1,50	1,75	2,25

6. Найти уравнение регрессии Y на X по данным, приведенным в корреляционной таблице:

$Y \backslash X$	20	25	30	35	40	Σ
16	4	6				10
26		8	10			18
36			32	3	9	44
46			4	12	6	22
56				1	5	6
Σ	4	14	46	16	20	100

Вариант 3

4. Произведено по четыре испытания на каждом из трех уровней фактора F . Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве средних для разных уровней фактора. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Результаты измерений приведены в таблице:

Номер испытания	Уровни фактора		
i	F ₁	F ₂	F ₃
1	35	30	21
2	32	24	22
3	31	26	34
4	30	20	31
$\Sigma/4$	32	25	27

Указание. Принять $y_{ij} = x_{ij} - 28$.

5. В таблице приведены данные об индексе розничных цен на пищевые товары и индексе промышленного производства:

Год	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958
Индекс цен X	100	101	113	115	113	113	111	112	115	120
Индекс производства Y	64	75	81	91	91	85	96	99	100	93

Написать формулу прогноза среднего индекса производства по цене X.

6. Найти уравнение регрессии Y на X и X на Y по данным, приведенным в корреляционной таблице:

YX	5	10	15	20	25	30	35	40	Σ
100	2	1							3
120	3	4	3						10
140			5	10	8				23
160				1		6	1	1	9
180							4	1	5
Σ	5	5	8	11	8	6	5	2	50

Вариант 4

4. Произведено по семь испытаний на каждом из четырех уровней фактора F . Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве средних для разных уровней фактора. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Результаты измерений приведены в таблице:

Номер испытания	Уровни фактора			
i	F_1	F_2	F_3	F_4
1	51	52	56	54
2	59	58	56	58
3	53	66	58	62
4	59	69	58	64
5	63	70	70	66
6	69	72	74	67
7	72	74	78	69
$\Sigma/7$	60,9	65,9	64,3	62,9

Указание. Принять $y_{ij} = x_{ij} - 63$.

5. Имеются следующие ряды оценок по тестам чтения и арифметики:

Чтение	43	58	45	53	37	58	55	61	46	64	46	62	60	56
Арифметика	32	25	28	30	22	25	22	20	20	30	21	28	34	28

Вычислите коэффициент корреляции.

6. Найти уравнение регрессии Y на X и X на Y по данным, приведенным в корреляционной таблице:

$Y \backslash X$	18	23	28	33	38	43	48	Σ
125		1						1
150	1	2	5					8
175		3	2	12				17
200			1	8	7			16
225					3	3		6
250						1	1	2
Σ	1	6	8	20	10	4	1	50

Вариант 5

4. Произведено по четыре испытания на каждом из трех уровней фактора F. Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве средних для разных уровней фактора. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Результаты измерений приведены в таблице:

Номер испытания	Уровни фактора		
	F_1	F_2	F_3
1	27	24	22
2	23	20	21
3	29	26	36
4	29	30	37
$\Sigma/4$	27	25	29

Указание. Принять $y_{ij} = x_{ij} - 27$.

5. Имеются следующие данные по десяти шахтам о сменной добыче угля на одного рабочего $Y(t)$ и мощности пласта $X(m)$:

x_i	8	11	12	9	8	8	9	9	8	12
y_i	5	10	10	7	5	6	6	5	6	8

Оценить сменную среднюю добычу угля на одного рабочего для шахт с мощностью пласта 8 м, доверительный интервал для среднего и индивидуального значения для этой мощности.

6. Найти уравнение регрессии Y на X и X на Y по данным, приведенным в корреляционной таблице:

YX	5	10	15	20	25	30	35	Σ
100						6	1	7
120						4	2	6
140			8	10	5			23
160	3	4	3					10
180	2	1		1				4
Σ	5	5	11	11	5	10	3	50

ТРЕНИНГ УМЕНИЙ

1. Пример выполнения упражнения тренинга на умение № 1

Задание

Выполнение норм выработки рабочих характеризуется данными из таблицы:

Процент выполнения	90-100	100-110	110-120	120-130	130-140	Σ
Середина интервала	95	105	115	125	135	
Число рабочих	10	160	100	60	20	350

Найти смещенную и несмещенную оценку для дисперсии выполнения норм выработки и 95%-й доверительный интервал для генерального среднеквадратического отклонения. Проверить гипотезу о том, что среднеквадратическое отклонение выполнения плана равно 10 (уровень значимости 0,05).

Решение

Заполните таблицу, подобрав каждому алгоритму конкретное содержание.

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
1.	Вычислить точечную оценку для среднего по формулам: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ — для выборки, заданной вариационным рядом; $\bar{x} = \frac{1}{n} \sum_{i=1}^k x_j m_j = \sum_{j=1}^k x_j \tilde{p}_j$ — для выборки, заданной таблицей группировки.	Данные заданы таблицей группировки, поэтому пользуемся второй формулой: $\bar{x} = 1/350(95 \cdot 10 + 105 \cdot 160 + 115 \cdot 100 + 125 \cdot 60 + 135 \cdot 20) = 39450/350 = 112,71.$
2.	Вычислить смещенную точечную оценку для дисперсии по формулам:	$S^2 = 1/350(95 \cdot 95 \cdot 10 + 105 \cdot 105 \cdot 160 + 115 \cdot 115 \cdot 100 + 125 \cdot 125 \cdot 60 + 135 \cdot 135 \cdot 20) - (112,71)^2 = 4478750/350 - 12703,544$

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
2.	$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$ – для выборки, заданной вариационным рядом; $S^2 = \frac{1}{n} \sum_{j=1}^k x_j^2 m_j - \bar{x}^2 = \sum_{j=1}^k (x_j - \bar{x})^2 \frac{m_j}{n}$ – для выборки, заданной таблицей.	$S^2 = 12796,428 - 12703,544 = 92,884.$
3.	Вычислить несмещенную точечную оценку для дисперсии по формуле: $s^2 = \frac{n}{n-1} S^2;$ $\sigma = \sqrt{s^2}$ – оценка для среднеквадратического отклонения.	$s^2 = 350/349 - 92,884 = 93,15;$ $\sigma = \sqrt{93,15} = 9,65.$
4.	Вычислить границы доверительного интервала для генеральной дисперсии, используя таблицу 1 приложения 2.	В таблице 1 для $n=349$ и $\beta=0,9$ с помощью линейной интерполяции находим коэффициенты для вычисления верхней и нижней границы доверительного интервала: $k_1 = 0,660 + 349 \cdot (1000 - 39) \cdot (0,918 - 0,660) = 0,75;$ $k_2 = 1,618 + 349 \cdot (1000 - 39) \cdot (1,094 - 1,618) = 1,43;$ $0,75 \cdot 93,15 < \sigma^2 < 1,43 \cdot 93,15;$ $69,86 < \sigma^2 < 132,99;$ $8,36 < \sigma < 11,53.$
5.	Проверка гипотезы о числовом значении дисперсии	Так как значение 10 покрывается построенным доверительным интервалом, гипотеза о том, что $\sigma=10$, не противоречит данным опыта.

Решите самостоятельно:

Задание 1.1

Распределение рабочих по времени, затраченному на обработку одной детали, дано в таблице:

Время обработки, мин	2-4	4-6	6-8	8-10	10-12	Σ
Середина интервала	3	5	7	9	11	
Число рабочих	42	73	154	205	26	500

Найти смещенную и несмещенную оценку для дисперсии времени обработки детали рабочим и 95%-й доверительный интервал для генерального среднеквадратического отклонения.

2. Пример выполнения упражнения тренинга на умение № 2

Задание

Расход сырья на единицу продукции по старой технологии составил:

Расход сырья x_i	304	307	308	Σ
Число изделий m_i	1	4	4	9

А по новой технологии:

Расход сырья x_i	303	304	306	308	Σ
Число изделий m_i	2	6	4	1	13

Для проверки гипотезы о совпадении генеральных средних для старой и новой технологии вычислили эмпирические значения $\bar{x}=307,11$, $\bar{y}=304,77$, $s_x^2=2,378$ и $s_y^2=1,555$. При уровне значимости 0,1 выполнить проверку о совпадении генеральных дисперсий (о том, что различие выборочных оценок дисперсий незначимо) для того, чтобы проверить гипотезу о совпадении средних по критерию Стьюдента.

Решение

Заполните таблицу, подобрав каждому алгоритму конкретное содержание.

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
1.	Вычислить несмещенные оценки для генеральных дисперсий обеих совокупностей s_1^2 и s_2^2 (см. умение №1) и статистику для проверки гипотезы – их отношение $F = s_1^2/s_2^2$.	$s_1^2 = s_x^2 = 2,378$; $s_2^2 = s_y^2 = 1,555$; $F = s_1^2/s_2^2 = 2,378/1,555 = 1,53$; $n_1 = 9 - 1 = 8$; $n_2 = 13 - 1 = 12$.
2.	Выполнить проверку гипотезы. Критерии проверки для гипотезы H_0 , состоящей в том,	В нашем случае $H_1: \sigma^2 = \sigma_1^2 \neq \sigma_2^2$, следовательно, надо выполнить проверки: $F > F_{\alpha/2, n_1-1, n_2-1}$ либо $F < F_{1-\alpha/2, n_1-1, n_2-1}$

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
	что $\sigma_1^2 = \sigma_2^2$, по статистике $F = \frac{s_1^2}{s_2^2}$ для уровня значимости α таковы: $H_1: \sigma^2 = \sigma_1^2 > \sigma_2^2 - F > F_{\alpha, n_1-1, n_2-1}$ $H_1: \sigma^2 = \sigma_1^2 < \sigma_2^2 - F > F_{\alpha/2, n_1-1, n_2-1}$ либо $F < F_{1-\alpha/2, n_1-1, n_2-1}$ $(F < \frac{1}{F_{\alpha/2, n_2-1, n_1-1}})$ Граница критической области F_{α, n_1-1, n_2-1} берется из таблицы критических точек распределения F (таблица 3 приложения 2). Для отыскания значений $F_{1-\alpha/2, n_1-1, n_2-1}$ надо воспользоваться равенством: для $P > 0,5 \quad F_P(k_1, k_2) = \frac{1}{F_{1-P}(k_2, k_1)}$ (очевидным, если выборы поменять местами).	$(F < \frac{1}{F_{\alpha/2, n_2-1, n_1-1}});$ $F_{\alpha/2, n_1-1, n_2-1} = 2,85;$ $F_{\alpha/2, n_2-1, n_1-1} = 3,28;$ $F_{1-\alpha/2, n_1-1, n_2-1} = 1 / F_{\alpha/2, n_2-1, n_1-1} = 1/3,28 = 0,3$. Следовательно, критическая область такова: $(0; 0,3); (2,85; \infty)$. Значение статистики 1,53 не попало в критическую область, следовательно, гипотезу о равенстве дисперсий расхода сырья по старой и новой технологии принимаем и проверку о равенстве математических ожиданий проводить можно.

Решите самостоятельно:

Задание 2.1

Сравниваются два способа обработки деталей. Первым способом обработано 12 деталей, вторым – 10. Выборочные дисперсии соответственно равны 39 и 26 единиц. На уровне значимости 0,1 выяснить, можно ли считать, что способы обработки деталей обладают существенно различными дисперсиями (дисперсия – это точность обработки).

$$n_1 = 12 \quad s_1^2 = 39 \quad \alpha = 0,1$$

$$n_2 = 10 \quad s_2^2 = 26 \quad H_0: \sigma_1 = \sigma_2$$

$$1) F = \frac{s_1^2}{s_2^2} = \frac{39}{26} = 1,5$$

$$3) F_{\frac{\alpha}{2}, k_1, k_2} = F_{0,05, 11, 9} = 3,1$$

$$2) k_1 = n_1 - 1 = 12 - 1 = 11$$

$$k_2 = n_2 - 1 = 10 - 1 = 9.$$

$$F_{\frac{\alpha}{2}, k_2, k_1} = F_{0,05; 9, 11} = 4,9.$$

$$4) \text{критич. обл.} \quad F > 3,1 \quad \text{и} \quad F < \frac{1}{4,9} = 0,204$$

5.) проверим. $1,5 > 3,1$
или
 $1,5 < 0,34$.

Но не попадает, т.е. проходит.

Ответ: точности обработки ориент.

Задание 2.2

С помощью критерия Бартлета выполнить ту же проверку для четырех способов обработки, если первым способом обработано 15 изделий, вторым – 20, третьим – 20, четвертым – 14 изделий и выборочные дисперсии для них соответственно равны 26, 39, 48 и 31 единица.

3. Примеры выполнения упражнений тренинга на умение № 3

Задание 3а

При 120 подбрасываниях игральной кости единица выпала 25 раз, двойка 19 раз, тройка 15 раз, четверка 22 раза, пятерка 15 раз, шестерка 21 раз. Согласуется ли это с тем, что игральная кость правильной формы?

Решение

Заполните таблицу, подобрав каждому алгоритму конкретное содержание.

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
1.	<i>Критерий согласия Пирсона (χ^2 – критерий).</i> Вычислить значение статистики U $U = \chi^2 = \sum_{i=1}^m \frac{(n_i - np_i)^2}{np_i},$ где n – число испытаний. В дискретном случае генеральные частоты p_i – это вероятности значений x_i и частоты n_i – сколько раз значение x_i встретилось в выборке. Если генеральное распределение непрерывное, то выборка группируется. Для интервалов группировки выборки (x_i, x_{i+1}) подсчитываются частоты n_i (число попаданий данных опыта в интервал) и теоретические вероятности p_i попадания в интервал (x_i, x_{i+1}). Для вычисления вероятностей p_i надо воспользоваться таблицами генерального распределения $F(x)$: $p_i = F(x_{i+1}) - F(x_i)$.	Для решения данной задачи следует воспользоваться критерием согласия Пирсона. Генеральное распределение – дискретное. Относительные частоты известны и равны: $p_i = 1/6$ ($i=1, 2, \dots, 6$); $n_1=25$; $n_2=19$; $n_3=15$; $n_4=22$; $n_5=15$; $n_6=21$; $n=120$; $np_i=120/6=20$. Вычисляем значение статистики U: $U = \chi^2 = \sum_{i=1}^m \frac{(n_i - np_i)^2}{np_i} =$ $= 1/20[(25-20)^2 + (19-20)^2 + (15-20)^2 + (22-20)^2 + (15-20)^2 + (21-20)^2] =$ $= 1/20(25+1+25+4+25+1) = 80/20 = 4.$

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
2.	Для выбранного уровня значимости α по таблице χ^2 распределения (таблица 2 приложения) найти критическое значение $\chi^2_{\alpha,k}$ для числа степеней свободы $k=m-r-1$ (m – число интервалов группировки или принимаемых значений, r – число параметров генерального распределения, замененных оценками).	Находим критическое значение $\chi^2_{\alpha,k}$ для числа степеней свободы $k=m-r-1=6-0-1=5$. $\chi^2_{0,05;5}=11,1$
3.	Сравнить U и $\chi^2_{\alpha,k}$. Если $U > \chi^2_{\alpha,k}$, то гипотеза H_0 отвергается; если $U < \chi^2_{\alpha,k}$, гипотеза H_0 не противоречит опытным данным.	Выясняем, попало ли значение статистики в критическую область: $4 < 11,1$, следовательно, в критическую область значение статистики не попало и гипотеза о симметричности игровой кости согласуется с данными опыта на уровне значимости 0,05.

Задание 36

Для анализа распределения диаметров валиков, изготовленных в течение двух разных смен, произведены две выборки. Размеры отклонений диаметра от стандартного занесены в таблицу:

Размер отклонения	Частоты выборки 1	Частоты выборки 2	Относит. частоты 1 $p_1/250$	Относит. частоты 2 $p_2/240$	Накопленные относит. частоты 1	Накопленные относит. частоты 2	Модуль разности
4,2	12	10	0,048	0,0417	0,048	0,0417	0,006
4,4	20	18	0,08	0,075	0,128	0,1167	0,011
4,6	22	20	0,088	0,083	0,216	0,1997	0,016
4,8	38	40	0,152	0,17	0,368	0,3517	0,016
5	86	80	0,344	0,33	0,712	0,6957	0,016
5,2	60	62	0,24	0,258	0,952	0,9357	0,016
5,4	8	4	0,032	0,017	0,984	0,9677	0,016
5,6	2	1	0,008	0,004	0,992	0,9717	0,02
5,8	1	2	0,004	0,008	0,996	0,9797	0,016
6,00	1	1	0,004	0,004	1	1	0
Σ	250	240					

При уровне значимости 0,05 проверить нулевую гипотезу о том, что распределение погрешностей обработки для обеих смен описывается одной и той же функцией распределения.

Решение

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
3.	<i>Проверка однородности двух выборок по критерию Колмогорова-Смирнова.</i> Сравнить две эмпирические функции распределения. Эмпирическая функция распределения – это накопленные частоты, деленные на число испытаний, т.е. накопленные относительные частоты. Статистика критерия Колмогорова-Смирнова имеет вид: $\lambda' = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot \max F_{n_1}(x) - F_{n_2}(x) ,$ где $F_{n_1}(x)$ и $F_{n_2}(x)$ – эмпирические функции распределения, построенные по двум выборкам объемов n_1 и n_2. Гипотеза H_0 отвергается, если фактически наблюдаемое значение статистики λ' больше $\lambda'_{кр}$, и принимается в противном случае. При $n_1 > 50$ и $n_2 > 50$ статистика λ' имеет практически распределение Колмогорова. При меньших n надо пользоваться специальными таблицами.	Вычисляем накопленные относительные частоты для обеих выборок, их разности и находим максимальную разность. Она равна 0,02 (см. заполненную таблицу) $\lambda' = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot \max F_{n_1}(x) - F_{n_2}(x) =$ $= \sqrt{\frac{250 \cdot 240}{250 + 240}} \cdot 0,02 = 11,07 \cdot 0,02 = 0,22.$ Так как $n_1 > 50$ и $n_2 > 50$, статистика λ' имеет практически распределение Колмогорова. Для уровня значимости 0,05 критическое значение равно 1,36. Эмпирическое значение статистики 0,22 не попало в критическую область, так как $0,22 < 1,36$. Следовательно, гипотеза H_0 не отвергается и можно считать, что данные опыта не противоречат предположению о том, что обе смены работают одинаково.

Решите самостоятельно:

Задание 3.1

818.

При 120 подбрасываниях игральной кости единица выпала 30 раз, двойка 29 раз, тройка 25 раз, четверка 24 раза, пятерка 5 раз, шестерка 7 раз. Согласуется ли это с тем, что игральная кость правильной формы?

Задание 3.2

В таблице представлены статистические данные о рождении четверень в год на протяжении 69 лет:

Число четверень	0	1	2	3	4	5	6	7	Σ
Частота	6	26	18	10	3	3	2	1	69

Проверить при уровне значимости 0,05 степень согласия с этими данными гипотезы о том, что количество рождений четверень в год подчиняется пуассоновскому закону распределения (*Указание.* В качестве λ закона Пуассона взять эмпирическую оценку для него – выборочное среднее, число степеней свободы уменьшить на 1).

Задание 3.3

Взять у двух студентов группы с сильно различающимися номерами фамилий таблицы группировки выборки из равномерного распределения или из нормального распределения и проверить гипотезу о том, что у обеих выборок одинаковые генеральные распределения.

4. Примеры выполнения упражнения тренинга на умение № 4

Задание

Произведено по четыре испытания на каждом из трех уровней фактора F. Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Результаты испытаний приведены в таблице:

Номер испытания	F ₁	F ₂	F ₃
1	38	20	21
2	36	24	22
3	35	26	31
4	31	30	34
$\Sigma/4$	35	25	27

Решение

Заполните таблицу, подобрав каждому алгоритму конкретное содержание.

Для упрощения расчетов можно перейти к данным: $y_{ij}=10 \cdot x_{ij}-C$. Вычтем из всех значений $29=(35+25+27)/3$ и составим новую расчетную таблицу, дополнив ее столбцами, содержащими квадраты исправленных выборочных значений и строками для хранения сумм по строке и столбцу:

$i \backslash j$	F_1		F_2		F_3		Σ
	y_{i1}	y_{i1}^2	y_{i2}	y_{i2}^2	y_{i3}	y_{i3}^2	
1	9	81	-9	81	-8	64	
2	7	49	-5	25	-7	49	
3	6	36	-3	9	2	4	
4	2	4	1	1	5	25	
$R_j = \sum_{i=1}^n y_{ij}^2$		170		116		142	$\Sigma R_j=428$
$T_j = \sum_{i=1}^n y_{ij}$	24		-16		-8		$\Sigma T_j=0$
T_j^2	576		256		64		$\Sigma T_j^2=896$

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
1.	Проверка выполняется средствами дисперсионного анализа. В дисперсионном анализе для каждого из j ($j=1,2,\dots,m$) уровней фактора проводится n испытаний Пусть x_{ij} – значение измеряемого признака X , полученное в i -м опыте при j -м уровне фактора. Можно перейти от x_{ij} к $y_{ij}=10 \cdot x_{ij}-C$. Результаты	Число уровней фактора $m=3$. Число испытаний на каждом уровне $n=4$. Вычисляем значения y_{ij}^2 и помещаем их в таблицу. Вычисляем значения R_j , T_j и T_j^2 и помещаем их в три последние строки таблицы.

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
	наблюдений располагаем в двумерной таблице размерности $n \cdot m$. Будем считать, что данные заданы так, что i (номер строки) – это номер наблюдения, а j (номер столбца) – это номер уровня фактора. Для каждого уровня фактора j (столбца) вычислить суммирование по столбцу: $R_j = \sum_{i=1}^n x_{ij}^2; T_j = \sum_{i=1}^n x_{ij}$ или $R_j = \sum_{i=1}^n y_{ij}^2; T_j = \sum_{i=1}^n y_{ij}$, если мы перешли от x_{ij} к $y_{ij} = 10^k x_{ij} - C$.	
2.	Выполнить суммирование трех последних строк по строке.	Результаты суммирования помещаем в последнем столбце трех последних строк.
3.	Вычислить Q, Q_1 и Q_2, используя результаты суммирования последних трех строк по строке: $Q = \sum_{j=1}^m R_j - \frac{1}{mn} \left(\sum_{j=1}^m T_j \right)^2$ $Q_1 = \frac{1}{n} \sum_{j=1}^m T_j^2 - \frac{1}{mn} \left(\sum_{j=1}^m T_j \right)^2$ $Q_{\text{ост}} = Q_{\text{общ}} - Q_{\text{факт}}$ или $Q_2 = Q - Q_1$	Число уровней фактора $m=3$. Число испытаний на каждом уровне $n=4$. $Q_{\text{общ}} = \sum R_j - 1/3 \cdot 4 \cdot [\sum T_j]^2 = 428 - 0 = 428$. $Q_{\text{факт}} = Q_1 = [\sum T_j^2] / 4 - 1/3 \cdot 4 \cdot [\sum T_j]^2 = 896/4 - 0 = 224$. $Q_{\text{ост}} = Q_2 = Q_{\text{общ}} - Q_{\text{факт}} = 428 - 224 = 204$.
4.	Найти несмещенные оценки факторной и остаточной дисперсий. Для этого разделить Q_1 и Q_2 на соответствующее число	$s_{\text{факт}}^2 = s_{\text{факт}}^2 = Q_1 / (m-1) = 224 / (3-1) = 112$; $s_{\text{ост}}^2 = s_{\text{ост}}^2 = Q_2 / (mn-m) = 204 / (12-3) = 22,67$.

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
	степеней свободы. Число степеней свободы – это число квадратов в сумме минус число связывающих их уравнений. Поэтому для $s^2_{\text{факт}} = s^2_1$ число степеней свободы равно $m-1$, а для $s^2_{\text{ост}} = s^2_2 - mn - m$. Таким образом, $s^2_1 = Q_1/(m-1)$, а $s^2_2 = Q_2/(mn-m)$.	
5.	Вычислить значение статистики $F = \frac{s^2_1}{s^2_2}$. Найти в таблице 3 приложения 2 критическое значение $F_{\alpha; k_1; k_2}$.	$F = \frac{s^2_1}{s^2_2} = 112/22,67 = 4,94;$ $F_{0,05; 2; 9} = 4,26.$
6.	Принять решение о непротиворечивости гипотезы H_0 . Гипотеза H_0 отвергается, если фактически вычисленное значение статистики $F > F_{\alpha; k_1; k_2}$, определенного на уровне значимости α при числе степеней свободы $k_1 = m-1$, $k_2 = mn-m$, и принимается, если $F < F_{\alpha; k_1; k_2}$. Опровержение гипотезы означает наличие существенных различий в зависимости от уровня фактора. Критическое значение $F_{\alpha; k_1; k_2}$ отыскивается в таблице 3 приложения 2.	$4,94 > 4,26.$ Значение статистики попало в критическую область, нулевую гипотезу о равенстве средних для разных уровней фактора отвергаем, они различаются значимо.
7.	Вычислить оценки для генерального среднего, если H_0 проходит, или генеральных средних, если H_0 не проходит. Если гипотеза H_0 о равенстве математических ожиданий при разных уровнях фактора не отвергается, т.е. влияние фак-	Оценки параметров: $\bar{y} = 0; \bar{x} = 29;$ $\lambda_1 = 35 - 29 = 6;$ $\lambda_1 = 25 - 29 = -4;$ $\lambda_1 = 27 - 29 = -2;$ $s^2_{\text{ост}} = 22,67$ – оценка для остаточной дисперсии.

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложен- ному алгоритму
	тора А не подтверждается, то эффекты влияния уровней фактора λ_j будут нулевыми. Остальные параметры по результатам наблюдений оцениваются следующим образом: - оценка генерального среднего $\bar{x}$ равна общему среднему по всем наблюдениям $\bar{x} = \frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n x_{ij} = \frac{1}{mn} \sum_{j=1}^m T_j$ (в случае перехода к переменным $y_{ij} = 10^k x_{ij} - C$ следует не забыть сделать соответствующее преобразование, например, если $y_{ij} = x_{ij} - C$, к этому значению надо добавить C, т.е. в этом случае $\begin{aligned} \bar{x} &= \frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n x_{ij} = \frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n (y_{ij} + C) = \\ &= C + \frac{1}{mn} \sum_{j=1}^m T_j; \end{aligned}$ - оценка для σ^2: $s_{\text{ост}}^2 = \frac{1}{mn - m} \sum_{j=1}^m \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2.$ Если же гипотеза H_0 отвергается, т.е. влияние фактора не подтверждается по результатам наблюдений, то оценки для λ_j уже будут не нулевыми, а равными разности $\bar{x}_j - \bar{x}$, где $\bar{x}_j$ – групповое среднее на j-м уровне фактора А.	
8.	Вычислить коэффициент детерминации – отношение факторной вариации к общей: $\rho^2 = Q_{\text{факт}} / Q_{\text{общ.}}$ Выписать в словесной форме, что доля “ρ^2 дисперсии признака Х объясняется зависимостью от фактора”.	$\rho^2 = 224/428 = 0,523$, т.е. примерно половина (0,523 часть) дисперсии признака Х объясняется зависимостью от фактора.

Решите самостоятельно:**Задание 4.1**

Имеются четыре партии сырья для текстильной промышленности. Из каждой партии отобрано по пять образцов и проведены испытания на определение величины разрывной нагрузки. Результаты испытаний приведены в таблице.

Номер партии	Разрывная нагрузка, кг/см ²				
1	200	140	170	145	165
2	190	150	210	150	150
3	230	190	200	190	200
4	150	170	150	170	180

Необходимо выяснить, существенно ли влияние различных партий сырья на величину разрывной нагрузки.

Указание. Фактором является партия, так что стоит записать результаты испытаний в транспонированной форме:

Номер измерения	Номер партии (фактор)			
	1	2	3	4
1	200	190	230	150
2	140	150	190	170
3	170	210	200	150
4	145	150	190	170
5	165	150	200	180
$\Sigma/5$	164	170	202	164

$$\bar{x} = 1/4(164+170+202+164)=175.$$

Для упрощения счета можно перейти к $y_i = x_i - 175$.

5. Пример выполнения упражнения тренинга на умение № 5

Задание

Стоимость акций по годам некоторой фирмы указана в таблице:

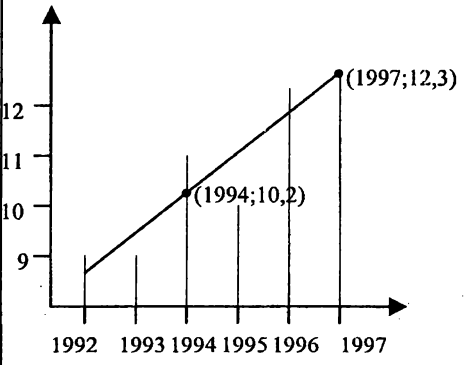
Год t_i	1992	1993	1994	1995	1996
Стоимость y_i	9	9	11	10	12

Для выявления тенденции стоимости на ближайший год найти методом наименьших квадратов по этим данным прямую линию, вычислить прогнозируемое значение стоимости акций на 1997 г. и построить для этого значения доверительный интервал.

Решение

Заполните таблицу, подобрав каждому алгоритму конкретное содержание.

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
1.	Составить таблицу для расчета. В таблице должны присутствовать столбцы: $\Delta x_i, y_i \Delta x_i, (\Delta x_i)^2, y_i^2, ax_i+b,$ $\Delta_i = y_i - ax_i - b, (\Delta_i)^2$ и две последние строки – для сумм по столбцу и этих сумм, деленных на n .	Поскольку начало отсчета времени можно выбрать произвольно, то для облегчения счета примем 1994 год и введем новую переменную $x = t - 1994$ $\begin{array}{cccccccc} \Delta x_i & y_i & y_i \Delta x_i & (\Delta x_i)^2 & y_i^2 & ax_i+b & \Delta_i & (\Delta_i)^2 \\ -2 & 9 & -18 & 4 & 81 & 8,8 & 0,2 & 0,04 \\ -1 & 9 & -9 & 1 & 81 & 9,5 & -0,5 & 0,25 \\ 0 & 11 & 0 & 0 & 121 & 10,2 & 0,8 & 0,64 \\ 1 & 10 & 10 & 1 & 100 & 10,9 & -0,9 & 0,81 \\ 2 & 12 & 24 & 4 & 144 & 11,6 & 0,4 & 0,16 \\ \hline \Sigma & 0 & 51 & 7 & 10 & 527 & & 1,9 \\ \Sigma/n & 0 & 10,2 & 1,4 & 2 & 105,4 & -(10,1)^2 & 3,39 \end{array}$ $\Delta_i = y_i - ax_i - b$
2.	Вычислить коэффициенты прямой. Уравнение прямой ищем в виде $y = a(x - \bar{x}) + \bar{y}$, где	Выполняем вычисления: $\bar{x} = 0; \bar{y} = 51/5 = 10,2;$ $a = (\sum y_i \Delta x_i) / \sum (\Delta x_i)^2 = 7/10 = 0,7.$

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
	$a = \frac{\frac{1}{n} \sum x_i y_i - \frac{1}{n} \sum x_i \cdot \frac{1}{n} \sum y_i}{\frac{1}{n} \sum x_i^2 - \left(\frac{1}{n} \sum x_i \right)^2} =$ $= \frac{\sum y_i \Delta x_i}{\sum (\Delta x_i)^2};$ $\Delta x_i = x_i - \bar{x}; \quad b = \bar{y} - a \cdot \bar{x};$ $\bar{x} = \frac{1}{n} \sum x_i; \quad \bar{y} = \frac{1}{n} \sum y_i.$	Уравнение МНК – прямой: $y = 0,7x + 10,2$, где x измеряется в годах и отсчитывается от 1994 года.
3.	Построить график прямой [проводить прямую через две точки $(\bar{x}, \bar{y})$ и $(x_i, ax+b)$ с любым x].	1997 году соответствует $x=3$. Проводим прямую через две точки: $(\bar{x}, \bar{y}) = (1994; 10,2)$ $(1997; 12,3)$. Прогнозируемая стоимость акций на 1997 год $y = 0,7 \cdot 3 + 10,2 = 12,3$. 
4.	Вычислить коэффициент корреляции r. Для системы точек, заданной таблицей из двух строк, он вычисляется по формуле $r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} =$	Вычисление коэффициента корреляции r: $s_x^2 = 10/5 = 2; \quad s_y^2 = 3,39;$ $s_x = \sqrt{2} = 1,41; \quad s_y = \sqrt{3,39} = 1,84;$ $r = \frac{\frac{1}{n} \sum y_i \Delta x_i}{s_x s_y} = \frac{1,4}{1,41 \cdot 1,84} = 0,54.$

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
	$\frac{\frac{1}{n} \sum_{i=1}^n (x_i y_i) - \bar{x} \cdot \bar{y}}{\sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{y}^2}} =$ $\frac{\frac{1}{n} \sum y_i \Delta x_i}{s_x s_y}.$	
5.	Вычислить прогнозируемое значение и доверительный интервал для него. Доверительный интервал для прогнозируемого значения вычисляется по формуле: $ax+b-t_{1-\alpha;n-2} \cdot s_{Y(x)} < Y(x) < ax+b+t_{1-\alpha;n-2} \cdot s_{Y(x)},$ где $s_{Y(x)}^2 = S^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$ $S^2 = 1/(n-2) \sum_{i=1}^n (ax_i + b - y_i)^2 -$ оценка дисперсии σ^2	Прогнозируемое значение для 1997 года: $Y(1997)=0,7 \cdot 3 + 10,2 = 12,3.$ Оценка для дисперсии σ^2: $S^2 = 1/(n-2) \sum \Delta_i = 1/3 \sum (y_i - ax_i - b)^2 =$ $= 1/3 \cdot 1,9 = 0,63$ $S_{Y(x)}^2 = 0,63 [1 + 1/5 + (3 \cdot 3)/10] = 1,323$ $s_{Y(x)} = 1,15; t_{1-\alpha;n-2} = 3,18;$ $(\alpha = 0,05, n = 5, n - 2 = 3).$ $0,7 \cdot 3 + 10,2 - 3,18 \cdot 1,15 =$ $= 12,3 - 3,657 = 8,64$ $0,7 \cdot 3 + 10,2 + 3,18 \cdot 1,15 = 12,3 + 3,657 =$ $= 15,96$ Тем самым 95%-й доверительный интервал для стоимости акций в 1997 году: $8,64 < Y(1997) < 15,96.$
6.	Проверить значимость уравнения регрессии. Проверка значимости уравнения регрессии: $Q_R = \sum_{i=1}^n (ax_i + b - \bar{y})^2 - \text{сумма}$ квадратов отклонений, обусловленная регрессией, $Q_{\text{ост}} = \sum_{i=1}^n [y_i - (ax_i + b)]^2 -$	Вычисляем Q_R: $Q_R = 1,4 \cdot 1,4 + 0,7 \cdot 0,7 + 0 + 0,7 \cdot 0,7 +$ $+ 1,4 \cdot 1,4 = 4,9;$ $Q_{\text{ост}} = 1,9; F = 4,9 \cdot 3/1,9 = 7,74;$ $F_{\alpha;1;n-2} = 6,61.$ Так как $F > F_{\alpha;1;n-2}$ ($7,74 > 6,61$), уравнение регрессии значимо на уровне $\alpha = 0,05$. И то, что коэффициент корреляции не слишком близок к 1, и близость значения F к критическому

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
	остаточная сумма квадратов, характеризующая отклонение от среднего в точке x_i, т.е. влияние неучтенных факторов. Уравнение регрессии значимо на уровне α, если фактически наблюдаемое значение статистики $F = \frac{Q_R(n-m)}{Q_{\text{ост}}(m-1)} = \frac{Q_R(n-2)}{Q_{\text{ост}}(2-1)} > F_{\alpha;1;n-2},$ где $F_{\alpha;1;n-2}$ – табличное значение F-критерия Фишера-Снедекора, определенное на уровне значимости α; n – число наблюдений. Число m оцениваемых по выборке параметров в нашем случае равно двум (а и b). Значение F показывает, в какой мере регрессия лучше оценивает значение зависимой переменной по сравнению со средней (именно поэтому в данном случае проверяемая гипотеза проходит при попадании статистики F в критическую область).	значению показывают, что линейная зависимость от времени не слишком хорошо описывает тенденцию изменения стоимости акций, хотя в первом приближении эту связь можно считать статистической линейной.

Решите самостоятельно:

Задание 5.1

Готовая продукция фирмы за 6 лет соответственно составила (в тысячах рублей):

Год	1989	1990	1991	1992	1993
Продукция	18	21	26	22	25

Для выявления тенденции на ближайший год найти методом наименьших квадратов по этим данным прямую линию, вычислить прогнозируемую прибыль на следующий год и доверительный интервал для нее.

Задание 5.2

Для 10 партий товара имеем следующие значения размера x партии и затрат времени на ее производство y :

x_i	10	15	16	8	15	20	18	15	12	17
y_i	50	63	61	50	56	72	62	60	54	70

Сосчитать эмпирический коэффициент корреляции и на его основе сделать вывод о наличии и виде зависимости размера партии товара от затрат на ее производство. Построить методом наименьших квадратов прямую, изображающую эту зависимость. Вычислить среднее значение затрат для партии размера $x=13$.

6. Пример выполнения упражнения тренинга на умение № 6

Задание

Найти коэффициент линейной корреляции между признаками X и Y и написать уравнения прямых регрессии для корреляционной таблицы признаков:

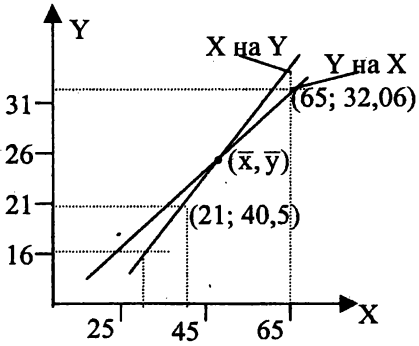
X\Y	11	16	21	26	31	36	$n_{i\cdot}$
25	2	4					6
35		6	3				9
45			6	45	4		55
55			2	8	6		16
65				4	7	3	14
$n_{\cdot j}$	2	10	11	57	17	3	100

Решение

Заполните таблицу, подобрав каждому алгоритму конкретное содержание.

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
1.	Вычислить (если они не заданы) в корреляционной таблице для каждого x_i значения $n_{i\cdot} = \sum_{j=1}^m n_{ij}$ (сколько раз такое значение x_i встретилось в паре с любыми y_j) и для каждого y_j значения $n_{\cdot j} = \sum_{i=1}^k n_{ij}$ (если X заданы по вертикали, а Y по горизонтали; если же X заданы по горизонтали, а Y по вертикали, то индексы просто меняются местами).	В нашей корреляционной таблице эти значения заданы
2.	Вычислить: $\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_{i\cdot}$ $s_x^2 = \frac{\sum_{i=1}^k x_i^2 n_{i\cdot}}{n} - (\bar{x})^2$	$\bar{x} = 1/100(25 \cdot 6 + 35 \cdot 9 + 45 \cdot 55 + 55 \cdot 16 + 65 \cdot 14) = 47,30;$ $s_x^2 = 1/100(25 \cdot 25 \cdot 6 + 35 \cdot 35 \cdot 9 + 45 \cdot 45 \cdot 55 + 55 \cdot 55 \cdot 16 + 65 \cdot 65 \cdot 14) - 47,3 \cdot 47,3 = 2337 - 2237,29 = 99,71; s_x = 9,985;$ $\bar{y} = 1/100(11 \cdot 2 + 16 \cdot 10 + 21 \cdot 11 + 26 \cdot 57 + 31 \cdot 17 + 36 \cdot 3) = 25,30;$

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
	$\bar{y} = \frac{1}{n} \sum_{j=1}^m y_j \cdot n_j, s_y^2 = \frac{\sum_{j=1}^m y_j^2 \cdot n_j}{n} - (\bar{y})^2$ $\mu = \overline{xy} - \bar{x} \cdot \bar{y} = \frac{\sum_{i=1}^k \sum_{j=1}^m x_i y_j \cdot n_{ij}}{n} - \bar{x} \cdot \bar{y}$ $r_{xy} = \frac{\mu}{s_x \cdot s_y}$ Значимость коэффициента корреляции (то, что генеральный коэффициент корреляции $\rho \neq 0$) проверяется с помощью статистики: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$ Гипотеза о том, что $\rho=0$, отвергается, если $ t = \frac{ r \sqrt{n-2}}{\sqrt{1-r^2}} > t_{1-\alpha; n-2},$ где $t_{1-\alpha; n-2}$ – табличное значение t-критерия Стьюдента, определенное на уровне значимости α при числе степеней свободы $n-2$.	$s_y^2 = 1/100(11 \cdot 11 \cdot 2 + 16 \cdot 16 \cdot 10 + 21 \cdot 21 \cdot 11 + 26 \cdot 26 \cdot 57 + 31 \cdot 31 \cdot 17 + 36 \cdot 36 \cdot 3) - 25,3 \cdot 25,3 = 664,10 - 640,09 = 24,01; s_y = 4,9;$ $\mu = 1/100(25 \cdot 11 \cdot 2 + 25 \cdot 16 \cdot 4 + 35 \cdot 16 \cdot 6 + 35 \cdot 21 \cdot 3 + 45 \cdot 21 \cdot 6 + 45 \cdot 26 \cdot 45 + 45 \cdot 31 \cdot 4 + 55 \cdot 21 \cdot 2 + 55 \cdot 26 \cdot 8 + 55 \cdot 31 \cdot 6 + 65 \cdot 26 \cdot 4 + 65 \cdot 31 \cdot 7 + 65 \cdot 36 \cdot 3) - 47,3 \cdot 25,3 = 1234,80 - 1196,69 = 38,11;$ $r = 38,11 / (9,985 \cdot 4,9) = 0,7789.$
3.	Выписать уравнения регрессии через коэффициенты a_{yx} и a_{xy} прямых регрессии, выписать тангенсы угла наклона прямых регрессии по формулам: $y_x - \bar{y} = a_{yx}(x - \bar{x}) = \frac{\mu}{s_x^2}(x - \bar{x});$ $x_y - \bar{x} = a_{xy}(y - \bar{y}) = \frac{\mu}{s_y^2}(y - \bar{y});$ $k_1 = a_{yx}; k_2 = \frac{1}{a_{xy}}.$	$a_{yx} = 38,11 / 99,71 = 0,382;$ $a_{xy} = 38,11 / 24,01 = 1,587;$ $k_1 = 0,392;$ $k_2 = 1 / 1,63 = 0,613.$ Прямые регрессии Y на X и X на Y: $y = 25,3 + 0,382(x - 47,30) = 0,382x + 7,23;$ $y(65) = 0,382 \cdot 65 + 7,23 = 32,06;$ $x = 47,30 + 1,587(y - 25,3) = 1,587y + 7,15;$ $x(21) = 1,587 \cdot 21 + 7,15 = 40,5;$ $x(16) = 1,587 \cdot 16 + 7,15 = 32,54.$

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
4.	Построить графики прямых регрессии [проводить прямые через точку $(\bar{x}, \bar{y})$ и еще одну точку, лежащую на прямой регрессии]. Сформулировать вывод о тесноте связи между компонентами.	 Близость к единице коэффициента корреляции, достаточно маленькая разница между углами наклона прямых регрессии говорит о тесной линейной связи компонент.

Решите самостоятельно:

Задание 6.1

При исследовании корреляционной зависимости по данным 20 предприятий между капиталовложениями X (млн. руб.) и выпуском продукции Y (млн. руб.) получены следующие уравнения регрессии: $y = 1,2x + 2$ и $x = 0,7y + 2$. Найти средние значения капиталовложений и выпуска продукции, а также коэффициент корреляции между рассматриваемыми признаками и оценить его значимость на 5%-ном уровне.

$$\begin{aligned}
 a_{yx} &= 1,2 & \bar{x} &= ? & r &= ? & a_{xy} &= \begin{cases} -1,2\bar{x} + \bar{y} = 2 \\ -0,7\bar{y} + \bar{x} = 2 \end{cases} \\
 a_{xy} &= 0,7 & \bar{y} &= ? & & & & \begin{cases} -0,84\bar{x} + 0,7\bar{y} = 1,4 \\ -0,7\bar{y} + \bar{x} = 2 \end{cases} \\
 r &= \pm \sqrt{a_{yx} \cdot a_{xy}} & & & & & & 0,16\bar{x} = 3,4 \\
 r &= \sqrt{1,2 \cdot 0,7} = \sqrt{0,84} = 0,9 & & & & & & \bar{x} = 21,25 \\
 & & & & & & & \bar{y} = 2 + 1,2 \cdot 21,25 = 27,5
 \end{aligned}$$

$$\begin{cases} y_x - \bar{y} = a_{yx}(x - \bar{x}) \\ x_y - \bar{x} = a_{xy}(y - \bar{y}) \end{cases}$$

$$\begin{cases} y_x - \bar{y} = 1,2(1,2x - 1,2\bar{x}) \\ x_y - \bar{x} = 0,7(0,7y - 0,7\bar{y}) \end{cases}$$

$$\begin{cases} y_x = 1,2x - 1,2\bar{x} + \bar{y} \\ x_y = 0,7y - 0,7\bar{y} + \bar{x} \end{cases}$$

Задание 6.2

При исследовании корреляционной зависимости между ценой на нефть X и индексом нефтяных компаний Y получены следующие данные: $\bar{x} = 16,2$ (ден. ед.); $\bar{y} = 4000$ (у.е.); $s_x^2 = 4$; $s_y^2 = 500$; $\mu = 40$. Составить уравнения регрессии. Используя соответствующее уравнение регрессии, найти среднюю величину индекса при цене на нефть 16,5 ден. ед.

Задание 6.3

При исследовании корреляционной зависимости между объемом продукции X (единиц) и ее себестоимостью Y (тыс. руб.) получено следующее уравнение регрессии Y по X : $Y_x = -0,0004x + 4,22$. Составить уравнение регрессии X по Y , если коэффициент корреляции между этими признаками оказался равным $-0,8$, а средний объем продукции $\bar{x} = 3000$ ед.

СПИСОК ФОРМУЛ

Эмпирическая дисперсия

Вычисление выборочной дисперсии:

	Вариационный ряд задан последовательностью	Задана таблица частот вариационного ряда	Задана таблица относительных частот вариационного ряда
Среднее значение выборки $\bar{X}$	$\frac{1}{n} \sum_{i=1}^n x_i$	$\frac{1}{n} \sum_{j=1}^k x_j m_j$	$\sum_{j=1}^k x_j \frac{m_j}{n}$
Дисперсия (разброс) выборки S^2	$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 =$ $= \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$	$\frac{1}{n} \sum_{j=1}^k x_j^2 m_j - \bar{x}^2 =$ $= \frac{1}{n} \sum_{j=1}^k (x_j - \bar{x})^2 m_j$	$\sum_{j=1}^k x_j^2 \frac{m_j}{n} - \bar{x}^2 =$ $= \sum_{j=1}^k (x_j - \bar{x})^2 \frac{m_j}{n}$

Несмещенная оценка для дисперсии $s^2 = \frac{n}{n-1} S^2$:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad - \text{ для вариационного ряда общего вида;}$$

$$s^2 = \frac{n}{n-1} \sum_{i=1}^k (x_i - \bar{x})^2 \frac{m_i}{n} \quad - \text{ для выборки, заданной таблицей.}$$

$$\text{Распределение выборочной дисперсии: } \frac{nS^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} = \chi_{n-1}^2,$$

следовательно, доверительный интервал ищется с помощью таблиц χ^2 распределения или с помощью таблицы 1 приложения 2.

Проверка гипотез о числовом значении дисперсии

Критерии проверки для гипотезы H_0 , состоящей в том, что $\sigma = \sigma_0$, по статистике χ^2 для уровня значимости α приведены в таблице:

Альтернативная гипотеза	Критерий отклонения гипотезы H_0 по статистике $\chi^2 = \frac{nS^2}{\sigma_0^2}$
$\sigma^2 = \sigma_1^2 > \sigma_0^2$	$\chi^2 > \chi^2_{\alpha, n-1}$
$\sigma^2 = \sigma_1^2 < \sigma_0^2$	$\chi^2 < \chi^2_{1-\alpha, n-1}$
$\sigma^2 = \sigma_1^2 \neq \sigma_0^2$	$\chi^2 > \chi^2_{\alpha/2, n-1}$ либо $\chi^2 < \chi^2_{1-\alpha/2, n-1}$

Граница критической области $\chi^2_{\alpha, k}$ берется из таблицы критических точек распределения χ^2 (таблица 2 приложения 2).

Проверка гипотез о совпадении дисперсий двух совокупностей

Критерии проверки гипотезы H_0 , состоящей в том, что $\sigma_1^2 = \sigma_2^2$, по статистике F на уровне значимости α приведены в таблице:

Альтернативная гипотеза	Критерий отклонения гипотезы H_0 по статистике $F = \frac{s_1^2}{s_2^2}$
$\sigma^2 = \sigma_1^2 > \sigma_2^2$	$F > F_{\alpha, n_1-1, n_2-1}$
$\sigma^2 = \sigma_1^2 \neq \sigma_2^2$	$F > F_{\alpha/2, n_1-1, n_2-1}$ либо $F < F_{1-\alpha/2, n_1-1, n_2-1}$ ($F < \frac{1}{F_{\alpha/2, n_2-1, n_1-1}}$)

Граница критической области F_{α, n_1-1, n_2-1} берется из таблицы критических точек распределения F (таблица 3 приложения 2). Для отыскания значений $F_{1-\alpha/2, n_1-1, n_2-1}$ надо воспользоваться равенством

для $P > 0,5$

$$F_p(k_1, k_2) = \frac{1}{F_{1-P}(k_2, k_1)}$$

(очевидным, если выборки поменять местами).

Минимальное число опытов для сравнения двух дисперсий, при котором будут соблюдены требования к величине ошибок первого и второго рода, определяется с помощью формулы: надо найти такое n , при котором выполняется $F_{\alpha}^2(n, n) = \Theta$, где через Θ обозначено аль-

тернативное единице значение отношения $\frac{\sigma_1^2}{\sigma_2^2} = \Theta$ (для гипотезы H_0

это отношение равно единице, для гипотезы H_1 это отношение равно Θ). Такое n отыскивается на диагонали таблицы F-распределения. Минимальное число опытов равно $n+1$.

Проверка гипотез о законе распределения

Критерий согласия Пирсона (χ^2 -критерий)

В качестве меры расхождения U берется величина χ^2 , которая вычисляется следующим образом:

$$U = \chi^2 = \sum_{i=1}^m \frac{(n_i - np_i)^2}{np_i},$$

где n – объем выборки.

Если генеральное распределение дискретно, то p_i – это вероятности принять значение x_i , а n_i – эмпирические частоты для значения x_i (сколько раз значение x_i встретилось в выборке). Если генеральное распределение непрерывно или различных принимаемых значений очень много вообще, то выборка группируется. Для интервалов группировки выборки (x_i, x_{i+1}) подсчитываются частоты n_i (число попаданий данных опыта в интервал) и теоретические вероятности p_i попадания в интервал (x_i, x_{i+1}) . Для вычисления вероятностей p_i надо воспользоваться таблицами генерального распределения: $p_i = F(x_{i+1}) - F(x_i)$.

Схема применения χ^2 -критерия следующая:

1. Вычисляется значение статистики U .
2. Для выбранного уровня значимости α по таблице χ^2 распределения находят критическое значение $\chi_{\alpha, k}^2$ для числа степеней свободы $k = m - r - 1$ (m – число слагаемых в сумме, r – число параметров, замененных эмпирическими значениями).

3. Если $U > \chi_{\alpha, k}^2$, то гипотеза H_0 отвергается; если $U < \chi_{\alpha, k}^2$, гипотеза H_0 не противоречит опытным данным.

Критерий Колмогорова

Для проверки гипотезы о виде распределения вероятностей кроме критерия χ^2 часто используют критерий Колмогорова, в котором при вычислении меры расхождения между теоретическим и эмпирическим распределениями рассматривают максимальное значение абсолютной величины разности между эмпирической функцией распределения $F_n(x)$ и соответствующей теоретической функцией распределения $F(x)$:

$$D = \max |F_n(x) - F(x)|.$$

Здесь D – случайная величина. При неограниченном увеличении числа наблюдений ($n \rightarrow \infty$)

$$P\left\{\max_x |F_n(x) - F(x)| < \frac{t}{\sqrt{n}}\right\} \xrightarrow{n \rightarrow \infty} K(t),$$

где $K(t)$ – так называемое распределение Колмогорова. Взяв, например, $t=1,36$, из таблиц функции Колмогорова (таблица 10 приложения 2) получаем, что при достаточно большом объеме выборки неравенство

$|F_n(x) - F(x)| < \frac{1,36}{\sqrt{n}}$ (или $\sqrt{n} |F_n(x) - F(x)| < 1,36$) выполняется с вероятностью 0,95 (т.е. в среднем для 95% выборок). Таким образом, если взять величину $\lambda = \sqrt{n} |F_n(x) - F(x)|$ в качестве меры расхождения между теоретическим и эмпирическим распределением (она называется статистикой критерия Колмогорова), можно использовать то, что при больших n она имеет распределение Колмогорова.

Проверка однородности двух выборок по критерию Колмогорова-Смирнова

По критерию Колмогорова-Смирнова сравниваются две эмпирические функции распределения.

Статистика критерия Колмогорова-Смирнова имеет вид

$$\lambda' = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot \max |F_{n_1}(x) - F_{n_2}(x)|,$$

где $F_{n_1}(x)$ и $F_{n_2}(x)$ – эмпирические функции распределения, построенные по двум выборкам объемов n_1 и n_2 .

Гипотеза H_0 отвергается, если фактически наблюдаемое значение статистики λ' больше $\lambda'_{кр}$, и принимается в противном случае. При $n_1 > 50$ и $n_2 > 50$ статистика λ' имеет практически распределение Колмогорова. При меньших n надо пользоваться специальными таблицами.

Проверка однородности k выборок по критерию χ^2

Для проверки гипотезы об однородности k выборок надо вычислить значение статистики:

$$\chi_n^2 = \sum_{j=1}^k \sum_{i=1}^m \frac{(n_{ij} - n_j \cdot p_i)^2}{n_j \cdot p_i} = n \sum_{j=1}^k \sum_{i=1}^m \frac{\left(n_{ij} - 1/n \cdot \left(\sum_{j=1}^k n_{ij} \cdot \sum_{i=1}^m n_{ij} \right) \right)^2}{\sum_{j=1}^k n_{ij} \cdot \sum_{i=1}^m n_{ij}},$$

где $p_i = \sum_{j=1}^k n_{ij}$; $n = \sum_{j=1}^k \sum_{i=1}^m n_{ij}$,

и воспользоваться таблицей критических точек распределения χ^2 для числа степеней свободы $(m-1)(k-1)$.

Проверка независимости

Если признаки A и B независимы друг от друга и численность выборки n большая, то величина

$$\chi^2 = n \sum_{i=1}^r \sum_{j=1}^s \frac{\left(n_{ij} - \frac{a_i b_j}{n} \right)^2}{a_i b_j}, \text{ в которой}$$

$$a_i = \sum_{j=1}^s n_{ij}$$

$$b_j = \sum_{i=1}^r n_{ij}$$

имеет приближенно χ^2 -распределение с $(r-1)(s-1)$ степенями свободы. Критическая область для проверки гипотезы, что признаки A и B независимы друг от друга, определена на уровне значимости α неравенством $\chi^2 > \chi^2_{\alpha}$ при $f=(r-1)(s-1)$ степенях свободы.

Дисперсионный анализ. Сравнение параметров более чем двух случайных величин

В дисперсионном анализе для каждого из j уровней фактора проводится n испытаний (для простоты будем считать, что для всех уровней число опытов одно и то же).

Обозначим через x_{ij} значение измеряемого признака X , полученное в i -м опыте при j -м уровне фактора. Результаты наблюдений сведены в таблицу:

$i \setminus j$	1	2	...	m
Ном. набл. \ фактор	A_1	A_2	...	A_m
1	x_{11}	x_{12}	...	x_{1m}
2	x_{21}	x_{22}	...	x_{2m}
...	...	...	...	...
n	x_{n1}	x_{n2}	...	x_{nm}
Выборочное среднее	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_m$
Выборочная дисперсия	S^2_1	S^2_2	...	S^2_m

Здесь $\bar{x}_j = 1/n(x_{1j} + x_{2j} + \dots + x_{nj})$ и $S^2_j = 1/n \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$ (выборочное среднее и выборочная дисперсия на j -м уровне фактора).

Обозначим через $\bar{x}$ общую среднюю:

$$\bar{x} = \frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n x_{ij} = \frac{1}{m} \sum_{i=1}^m \bar{x}_i.$$

В дисперсионном анализе исследуется сумма квадратов отклонений:

$$Q_{\text{общ}} = \sum_{j=1}^m \sum_{i=1}^n (x_{ij} - \bar{x})^2,$$

которая разбивается на два слагаемых:

$$Q_{\text{общ}} = Q_{\text{факт}} + Q_{\text{ост}} = Q_1 + Q_2,$$

где $Q_{\text{факт}} = n \sum_{j=1}^m (\bar{x}_j - \bar{x})^2$ – факторная (межгрупповая) сумма квадратов отклонений групповых средних от общей средней;

$Q_{\text{ост}} = \sum_{j=1}^m \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$ – остаточная (внутригрупповая) сумма квадратов отклонений от групповых средних.

Деленные на соответствующее число степеней свободы, они являются несмещенными оценками соответствующих дисперсий. Число степеней свободы – это число квадратов в сумме минус число связывающих их уравнений. Поэтому для $s^2_{\text{факт}} = s^2_1$ число степеней свободы равно $m-1$, а для $s^2_{\text{ост}} = s^2_2$ – $mn - m$.

Таким образом, $s^2_1 = Q_1 / (m-1)$, а $s^2_2 = Q_2 / (mn-m)$.

Проверка нулевой гипотезы о равенстве m математических ожиданий сводится к проверке существенности различия несмещенных выборочных оценок s_1^2 и s_2^2 . Она осуществляется с помощью

статистики $F = \frac{s_1^2}{s_2^2}$.

Гипотеза H_0 отвергается, если фактически вычисленное значение статистики $F > F_{\alpha; k_1; k_2}$, определенного на уровне значимости α при числе степеней свободы $k_1 = m-1$, $k_2 = mn-m$, и принимается, если $F < F_{\alpha; k_1; k_2}$. Опровержение гипотезы означает наличие существенных различий в зависимости от уровня фактора. Критическое значение $F_{\alpha; k_1; k_2}$ отыскивается в таблице 3 приложения 2.

Если гипотеза H_0 о равенстве математических ожиданий при разных уровнях фактора не отвергается, т.е. влияние фактора A не подтверждается, то эффекты влияния уровней фактора λ_j будут нулевыми, остальные параметры по результатам наблюдений оцениваются следующим образом.

Оценка математического ожидания μ равна общему среднему по всем наблюдениям:

$$\bar{x} = \frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n x_{ij}.$$

Оценка для σ^2 :

$$s_0^2 = \frac{1}{mn - m} \sum_{j=1}^m \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2.$$

Если же гипотеза H_0 отвергается, т.е. влияние фактора подтверждается по результатам наблюдений, то оценки для λ_j уже будут не нулевыми, а равными разности $\bar{x}_j - \bar{x}$, где $\bar{x}_j$ - групповое среднее на j -м уровне фактора А.

Коэффициент детерминации – это отношение факторной вариации к общей (т.е. коэффициент детерминации указывает, какая доля дисперсии признака X объясняется зависимостью признака X от фактора):

$$\rho^2 = Q_{\text{факт}} / Q_{\text{общ.}}$$

Ковариационный и регрессионный анализ

Ковариация и коэффициент корреляции

Для двумерного дискретного распределения ковариация вычисляется по следующей формуле:

$$K_{xy} = \sum \sum x_j y_i p_{ij} - \sum x_j p_{*j} \cdot \sum y_i p_{i*}.$$

Коэффициент корреляции – отношение ковариации к произведению среднеквадратических отклонений величин X и Y :

$$\rho_{xy} = K_{xy} / s_x s_y.$$

Коэффициент корреляции обладает следующими свойствами:

1. $-1 \leq \rho \leq 1$.
2. Если случайные величины независимы, то их коэффициент корреляции равен нулю. Обратное утверждение, вообще говоря, неверно; из того, что $\rho=0$, еще не следует независимость случайных величин.
3. Если у двух случайных величин $|\rho|=1$, то между этими случайными величинами существует линейная функциональная зависимость.

Изучение пар случайных величин с помощью выборок

Для двумерных величин выборочные данные представляются в виде таблицы.

Таблицы могут состоять из двух строк:

x_i	x_1	x_2	...	x_n
y_i	y_1	y_2	...	y_n

Если данных много, то их группируют. Если одно и то же значение X в выборке встретилось n_x раз, а значение Y – n_y раз, а пара (x, y) – n_{xy} раз, данные наблюдений (группированные) записывают в виде двумерной (состоит из m строк и k столбцов) таблицы, которую называют корреляционной. В правом столбце указывают, сколько раз встретились конкретные значения y , в нижней строке – сколько раз встретились конкретные значения x , в клетке, расположенной в правом нижнем углу, – сумму всех частот (общее число всех наблюдений $n = \sum n_x = \sum n_y$):

$\begin{matrix} Y \\ \backslash \\ X \end{matrix}$	x_1	x_2	...	x_k	$\sum_{j=1}^k$
y_1	n_{11}	n_{12}	...	n_{1k}	$n_{1*} = \sum n_{1j}$
y_2	n_{21}	n_{22}	...	n_{2k}	$n_{2*} = \sum n_{2j}$
...	...	...	...	...	...
y_m	n_{m1}	n_{m2}	...	n_{mk}	$n_{m*} = \sum n_{mj}$
$\sum_{i=1}^m$	$n_{*1} = \sum n_{i1}$	$n_{*2} = \sum n_{i2}$	...	$n_{*k} = \sum n_{ik}$	$n = \sum n_{*i} = \sum n_{j*}$

По корреляционной таблице распределения выборочные условные средние вычисляются по следующим формулам:

$$\bar{y}_x = M(Y/X = x_j) = 1/n_{*j} \sum_{i=1}^m y_i n_{ij},$$

$$\bar{x}_y = M(X/Y = y_i) = 1/n_{i*} \sum_{j=1}^k x_j n_{ij}.$$

Построение прямой регрессии методом наименьших квадратов для системы точек, заданной таблицей из двух строк:

x_i	x_1	x_2	...	x_n
y_i	y_1	y_2	...	y_n

Уравнение МНК-прямой можно написать в такой форме:

$$y = a(x - \bar{x}) + \bar{y} = ax + b,$$

$$\text{где } a = \frac{\frac{1}{n} \sum x_i y_i - \frac{1}{n} \sum x_i \cdot \frac{1}{n} \sum y_i}{\frac{1}{n} \sum x_i^2 - \left(\frac{1}{n} \sum x_i \right)^2} = \frac{\sum y_i \Delta x_i}{\sum (\Delta x_i)^2} = \frac{\sum y_i \Delta x_i}{\sum x_i \cdot \Delta x_i};$$

$$b = \bar{y} - a\bar{x}; \quad \bar{x} = \frac{1}{n} \sum x_i; \quad \bar{y} = \frac{1}{n} \sum y_i; \quad \Delta x_i = x_i - \bar{x}$$

(во всех суммах знак $\sum$ означает суммирование по всем точкам $\sum_{i=1}^n$).

Эта прямая проходит через точку $(\bar{x}, \bar{y})$, являющуюся центром тяжести данной системы точек.

Доверительный интервал для прогнозов индивидуальных значений $Y(x)$, прогнозируемых по системе точек (x_i, y_i) $i=1, 2, \dots, n$ с помощью МНК-прямой, определяется по формуле

$$ax + b - t_{1-\alpha; n-2} \cdot s_{Y(x)} < Y(x) < ax + b + t_{1-\alpha; n-2} \cdot s_{Y(x)},$$

$$\text{где } s_{Y(x)}^2 = s^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right];$$

$$s^2 = 1/(n-2) \sum_{i=1}^n (ax_i + b - y_i)^2 \quad - \text{оценка дисперсии } \sigma^2;$$

$t_{1-\alpha; n-2}$ - отыскивается по уровню значимости α в таблице распределения Стьюдента для $n-2$ степеней свободы (таблица 9 приложения 2).

Доверительный интервал для прогнозов средних значений $y(x)$ отличается от такового для $Y(x)$ только тем, что в формуле для $s^2_{y(x)}$ отсутствует слагаемое 1.

Проверка значимости уравнения регрессии

$$Q_R = \sum_{i=1}^n (ax_i + b - \bar{y})^2 \text{ — сумма квадратов отклонений, обуслов-$$

ленная регрессией;

$$Q_{\text{ост}} = \sum_{i=1}^n [y_i - (ax_i + b)]^2 \text{ — остаточная сумма квадратов, харак-}$$

теризующая отклонение от среднего в точке x_i , то есть влияние неучтенных факторов.

Схема дисперсионного анализа описывается таблицей:

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Средние квадраты
Регрессия	Q_R	$m - 1$	$s^2_R = Q_R / (m - 1)$
Остаточная	$Q_{\text{ост}}$	$n - m$	$s^2 = Q_{\text{ост}} / (n - m)$

Уравнение регрессии значимо на уровне α , если фактически наблюдаемое значение статистики

$$F = \frac{Q_R (n - m)}{Q_{\text{ост}} (m - 1)} = \frac{Q_R (n - 2)}{Q_{\text{ост}} (2 - 1)} > F_{\alpha; 1; n-2},$$

где $F_{\alpha; 1; n-2}$ — табличное значение F-критерия Фишера-Снедекора, определенное на уровне значимости α ; n — число наблюдений. Число m оцениваемых по выборке параметров в нашем случае равно двум (a и b). Значение F показывает, в какой мере регрессия лучше оценивает значение зависимой переменной по сравнению с ее средней (именно поэтому в данном случае проверяемая гипотеза проходит при попадании статистики F в критическую область).

Эмпирический коэффициент корреляции r

Для системы точек, заданной таблицей из двух строк:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i y_i) - \bar{x} \cdot \bar{y}}{\sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{y}^2}}.$$

Для выборки, заданной двумерной таблицей размерности $m \cdot k$:

$$r_{xy} = \frac{\mu}{s_x \cdot s_y},$$

где величины μ , s_x и s_y надо вычислять по формулам, в которых учитываются кратность значений:

$$\mu = \overline{xy} - \bar{x} \cdot \bar{y} = \frac{\sum_{i=1}^k \sum_{j=1}^m x_i y_j n_{ij}}{n} - \bar{x} \cdot \bar{y} \quad - \text{выборочная ковариация};$$

$$s_x^2 = \frac{\sum_{i=1}^k x_i^2 n_{i*}}{n} - (\bar{x})^2 \quad - \text{выборочная дисперсия переменной } X,$$

$$\text{где } n_{i*} = \sum_{j=1}^m n_{ij}, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_{i*}$$

$$s_y^2 = \frac{\sum_{j=1}^m y_j^2 n_{*j}}{n} - (\bar{y})^2 \quad - \text{выборочная дисперсия переменной } Y,$$

$$\text{где } n_{*j} = \sum_{i=1}^k n_{ij}, \quad \bar{y} = \frac{1}{n} \sum_{j=1}^m y_j n_{*j}$$

При $n > 50$ для интервальной оценки генерального коэффициента ρ , для которого выборочный коэффициент r является точечной оценкой, интервальная оценка задается формулой:

$$r - k_{1-\alpha} \frac{\sqrt{1-r^2}}{\sqrt{n-2}} \leq \rho \leq r + k_{1-\alpha} \frac{\sqrt{1-r^2}}{\sqrt{n-2}},$$

где $k_{1-\alpha}$ ищется по таблицам нормального распределения (таблица 8 приложения 2).

Значимость коэффициента корреляции (то, что генеральный коэффициент корреляции $\rho \neq 0$) проверяется с помощью статистики: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.

Гипотеза о том, что $\rho=0$, отвергается, если $|t| = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} > t_{1-\alpha; n-2}$, где

$t_{1-\alpha; n-2}$ – табличное значение t-критерия Стьюдента, определенное на уровне значимости α при числе степеней свободы $n-2$.

Построение двух прямых регрессии (X и Y случайны и равноправны)

Уравнения прямых регрессии:

$$y_x - \bar{y} = a_{yx} (x - \bar{x}) = r \frac{s_y}{s_x} (x - \bar{x});$$

$$x_y - \bar{x} = a_{xy} (y - \bar{y}) = r \frac{s_x}{s_y} (y - \bar{y});$$

$$r = a_{yx} \frac{s_x}{s_y} = a_{xy} \frac{s_y}{s_x} = \pm \sqrt{a_{yx} a_{xy}} \quad \text{т.е. коэффициент корреляции } r$$

переменных X и переменных Y есть средняя геометрическая коэффициентов регрессии, имеющая их знак.

Тангенс ϕ , где ϕ – угол между прямыми регрессии

Их угловые коэффициенты соответственно равны

$$k_1 = a_{yx}, \quad k_2 = 1/a_{xy}.$$

Воспользовавшись формулой аналитической геометрии, с помощью элементарных выкладок получаем

$$\operatorname{tg} \varphi = \frac{k_2 - k_1}{1 + k_1 k_2} = \frac{1 - a_{yx} a_{xy}}{a_{xy} + a_{yx}} = \frac{1 - r^2}{r} \cdot \frac{s_x s_y}{s_x^2 + s_y^2}.$$

Двумерное нормальное распределение

Случайная величина (X, Y) распределена по двумерному нормальному закону, если ее совместная плотность имеет вид:

$$p(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-a_x}{\sigma_x}\right)^2 - 2\rho\frac{x-a_x}{\sigma_x}\frac{y-a_y}{\sigma_y} + \left(\frac{y-a_y}{\sigma_y}\right)^2\right]}.$$

Таким образом, двумерный нормальный закон распределения определяется пятью параметрами: a_x , a_y , σ_x , σ_y , ρ . Параметры a_x , a_y выражают математические ожидания величин X и Y , параметры σ_x , σ_y — их среднеквадратические отклонения, ρ — коэффициент корреляции между случайными величинами X и Y . Условные математические ожидания и дисперсии, сосчитанные по этому закону распределения, выражаются с помощью следующих формул:

$$M_y(X) = M(X/Y = y) = a_x + \rho \frac{\sigma_x}{\sigma_y}(y - a_y); \quad D_y(X) = D(X/Y = y) = \sigma_x^2(1 - \rho^2);$$

$$M_x(Y) = M(Y/X = x) = a_y + \rho \frac{\sigma_y}{\sigma_x}(x - a_x); \quad D_x(Y) = D(Y/X = x) = \sigma_y^2(1 - \rho^2).$$

Из этих формул вытекает, что линии регрессии $M_y(X)$ и $M_x(Y)$ нормально распределенных случайных величин представляют собой прямые линии (регрессии Y по X и X по Y линейны). Условные дисперсии постоянны и не зависят от значения второй компоненты. Второе свойство называется равноизменчивостью условных нормальных распределений. Из вида плотности двумерного нормального распределения следует, что если $\rho=0$, то компоненты X и Y независимы.

ПРИЛОЖЕНИЕ 2

Таблица 1

Границы доверительных интервалов для дисперсии в зависимости от числа степеней свободы и надежности

Для получения несмещенных доверительных границ интервалов для оценки дисперсии σ^2 при нормальном распределении случайных величин в выборке находится несмещенная оценка дисперсии s^2 , затем эту оценку умножают на коэффициенты, взятые из таблицы.

Пример. Пусть по выборке $n=10$ при неизвестном среднем $s^2=2,87$. Здесь число степеней свободы $k=9$, а границы доверительного интервала надежности 0,90 таковы: нижняя – $0,498 \cdot 2,87 = 1,429$, а верхняя – $2,481 \cdot 2,87 = 7,120$.

Число степеней свободы n	Надежность					
	0,90		0,95		0,99	
	Нижняя граница	Верхняя граница	Нижняя граница	Верхняя граница	Нижняя граница	Верхняя граница
5	0,402	3,691	0,348	5,054	0,270	9,927
6	0,432	3,201	0,377	4,211	0,297	7,569
7	0,457	2,876	0,402	3,679	0,319	6,239
8	0,479	2,651	0,424	3,314	0,341	5,304
9	0,498	2,481	0,443	3,048	0,360	4,680
10	0,514	2,348	0,460	2,844	0,376	4,243
11	0,529	2,242	0,475	2,683	0,391	3,897
12	0,543	2,153	0,489	2,553	0,405	3,635
13	0,555	2,080	0,502	2,440	0,417	3,424
14	0,566	2,018	0,513	2,354	0,429	3,243
15	0,576	1,965	0,525	2,272	0,440	3,091
16	0,586	1,918	0,534	2,208	0,450	2,961
17	0,594	1,876	0,544	2,147	0,461	2,833
18	0,602	1,839	0,552	2,095	0,470	2,739
19	0,610	1,805	0,560	2,050	0,477	2,664
20	0,617	1,776	0,568	2,006	0,487	2,573
30	0,671	1,584	0,626	1,743	0,548	2,121
39	0,703	1,490	0,660	1,618	0,586	1,914
1000	0,931		0,918	1,094		

Таблица 2

Значения $\chi^2_{\alpha,k}$ критерия Пирсона

$k \backslash \alpha$	0,99	0,98	0,95	0,90	...	0,1	0,05	0,02	0,01
1	0,00	0,00	0,00	0,02		2,71	3,84	5,41	6,64
2	0,02	0,04	0,10	0,21		4,605	5,99	7,82	9,21
3	0,11	0,18	0,35	0,58		6,25	7,82	9,84	11,3
4	0,30	0,43	0,71	1,06		7,78	9,49	11,7	13,3
5	0,55	0,75	1,14	1,61		9,24	11,1	13,4	15,1
6	0,87	1,13	1,63	2,20		10,6	12,6	15,0	16,8
7	1,24	1,56	2,17	2,83		12,0	14,1	16,6	18,5
8	1,65	2,03	2,73	3,49		13,4	15,5	18,2	20,1
9	2,09	2,53	3,32	4,17		14,7	16,9	19,7	21,7
10	2,56	3,06	3,94	4,86		16,0	18,3	21,2	23,2
11	3,05	3,61	4,58	5,58		17,3	19,7	22,6	24,7
12	3,57	4,18	5,23	6,30		18,5	21,0	24,1	26,2
13	4,11	4,46	5,89	7,04		19,8	22,4	25,5	27,7
14	4,66	5,37	6,57	7,79		21,1	23,7	26,9	29,1
15	5,23	5,98	7,26	8,55		22,3	25,0	28,3	30,6
16	5,81	6,61	7,96	9,31		23,5	26,3	29,6	32,0
17	6,41	7,26	8,67	10,1		24,8	27,6	31,0	33,4
18	7,02	7,91	9,39	10,9		26,0	28,9	32,3	34,8
19	7,63	8,57	10,1	11,6		27,2	30,1	33,7	36,2
20	8,26	9,24	10,8	12,4		28,4	31,4	35,0	37,6
21	8,90	9,92	11,6	13,2		29,6	32,7	36,3	38,9
22	9,54	10,6	12,3	14,0		30,8	33,9	37,7	40,3
23	10,2	11,3	13,1	14,8		32,0	35,2	39,0	41,6
24	10,9	12,0	13,8	15,7		33,2	36,4	40,3	43,0
25	11,5	12,7	14,6	16,5		34,4	37,7	41,7	44,3
26	12,2	13,4	15,4	17,3		35,6	38,9	42,9	45,6
27	12,9	14,1	16,1	18,1		36,7	40,1	44,1	47,0
28	13,6	14,8	16,9	18,9		37,9	41,3	45,4	48,3
29	14,3	15,6	17,7	19,8		39,1	42,6	46,7	49,6
30	14,9	16,3	18,5	20,6		40,3	43,8	48,0	50,9

Критические значения для F-распределения для $\alpha=0,05$
(по вертикали k_2 – число степеней свободы знаменателя,
по горизонтали k_1 – число степеней свободы числителя)

$k_2 \backslash k_1$	1	2	3	4	5	6	7	8	9	10	11	12
1	161	200	216	225	230	234	237	239	241	242	243	244
2	18,51	19,0	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,40	19,41
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,76	8,74
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93	5,91
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70	4,68
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60	3,57
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,31	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10	3,07
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94	2,91
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82	2,79
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72	2,69
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,63	2,60
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56	2,53
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,51	2,48
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45	2,42
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,41	2,38

Таблица 4

$$\text{Значения функции } p(m) = \frac{a^m}{m!} \cdot e^{-a}$$

a \ m	0	1	2	3	4	5	6	7
0.1	0.90484	09048	00452	00015	00000	00000	00000	00000
0.2	0.81873	16375	01638	00109	00006	00000	00000	00000
0.3	0.74082	22225	03334	00333	00025	00002	00000	00000
0.4	0.67032	26813	05363	00715	00072	00006	00000	00000
0.5	0.69653	30327	07582	01264	00158	00016	00001	00000
0.6	0.54881	32929	09879	01976	00296	00036	00004	00000
0.7	0.49659	34761	12166	02839	00497	00070	00008	00001
0.8	0.44933	35946	14379	03834	00767	00123	00016	00002
0.9	0.40657	36591	16466	04940	01112	00200	00030	00004
1	0.36788	36788	18394	06131	01533	00307	00051	00007
2	0.3534	27067	27067	18045	09022	03609	01203	00344
3	0.4979	14936	22404	22404	16803	10082	05041	02160
4	0.1832	07326	14653	19537	19537	15629	10420	05954
5	0.0674	03369	08422	14037	17547	17547	14622	10445
6	0.0248	01487	04462	08924	13385	16062	16062	13768
7	0.0091	00638	02234	05213	09123	12772	14900	14900

Таблица 5

Значения функции $\Phi(x) = \frac{2}{\sqrt{2\pi}} \cdot \int_0^x e^{-\frac{t^2}{2}} dt$

x	0	1	2	3	4	5	6	7	8	9
0,0	0,000	0,0080	0,0160	0,0238	0,0319	0,0399	0,0478	0,0558	0,0638	0,0717
0,1	0797	0876	0955	1034	1113	1192	1271	1350	1428	1507
0,2	1585	1663	1741	1819	1897	1974	2051	2128	2205	2282
0,3	2358	2434	2510	2586	2661	2737	2812	2886	2960	3035
0,4	3108	3182	3255	3328	3401	3473	3545	3616	3688	3759
0,5	3829	3899	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4907	4971	5035	5098
0,7	5161	5223	5285	5346	5407	5467	5527	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
0,9	6319	6372	6424	6476	6528	6579	6629	6679	6729	6778
1,0	6827	6875	6923	6970	7017	7063	7109	7154	7199	7243
1,1	7287	7330	7373	7415	7457	7499	7540	7580	7620	7660
1,2	7699	7737	7775	7813	7850	7887	7923	7959	7994	8029
1,3	8064	8098	8132	8165	8198	8230	8262	8293	8324	8355
1,4	8385	8415	8444	8473	8501	8529	8557	8584	8611	8638
1,5	8664	8690	8715	8740	8764	8789	8812	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9090
1,7	9109	9127	9146	9164	9181	9189	9216	9233	9249	9265
1,8	9281	9297	9312	9327	9342	9357	9371	9385	9399	9412
1,9	9426	9439	9451	9464	9476	9488	9500	9512	9523	9534
2,0	9545	9556	9566	9576	9586	9596	9606	9616	9625	9634
2,1	9643	9651	9660	9668	9676	9684	9692	9700	9707	9715
2,2	9722	9729	9736	9743	9749	9756	9762	9768	9774	9780
2,3	9786	9791	9797	9802	9807	9812	9817	9822	9827	9832
2,4	9836	9841	9845	9849	9853	9857	9861	9865	9869	9872
2,5	9876	9879	9883	9886	9889	9892	9895	9898	9901	9904
2,6	9907	9910	9912	9915	9917	9920	9922	0024	9926	9928
2,7	9931	9933	9935	9937	9938	9940	9942	9944	9946	9947
2,8	9949	9951	9952	9953	9955	9956	9958	9959	9960	9961
2,9	9963	9964	9965	9966	9967	9968	9969	9970	9971	9972

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	
3,0	0,9973	3,4	0,9993	3,8	0,9999		
3,1	0,9981	3,5	0,9995	3,9	0,9999		
3,2	0,9986	3,6	0,9997	4,0	0,999936		
3,3	0,9990	3,7	0,9998	4,5	0,999994		

Таблица 6

Квантили нормального распределения $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u_p} e^{-\frac{t^2}{2}} dt = P$

P	u_p	P	u_p	P	u_p	P	u_p	P	u_p
0,0	$-\infty$	0,2	-0,842	0,4	-0,253	0,6	2,253	0,8	0,842
0,01	-2,326	0,21	-0,806	0,41	-0,228	0,61	0,279	0,81	0,878
0,02	-2,054	0,22	-0,772	0,42	-0,202	0,62	0,305	0,82	0,915
0,03	-1,881	0,23	-0,739	0,43	-0,176	0,63	0,332	0,83	0,954
0,04	-1,751	0,24	-0,706	0,44	-0,151	0,64	0,358	0,84	0,994
0,05	-1,645	0,25	-0,674	0,45	-0,126	0,65	0,385	0,85	1,036
0,06	-1,555	0,26	-0,643	0,46	-0,10	0,66	0,412	0,86	1,080
0,07	-1,476	0,27	-0,613	0,47	-0,075	0,67	0,440	0,87	1,126
0,08	-1,405	0,28	-0,583	0,48	-0,50	0,68	0,468	0,88	1,175
0,09	-1,341	0,29	-0,553	0,49	-0,025	0,69	0,496	0,89	1,227
0,1	-1,282	0,3	-0,524	0,5	0	0,7	0,524	0,9	1,282
0,11	-1,227	0,31	-0,496	0,51	0,025	0,71	0,553	0,91	1,341
0,12	-1,175	0,32	-0,468	0,52	0,050	0,72	0,583	0,92	1,405
0,13	-1,126	0,33	-0,440	0,53	0,075	0,73	0,613	0,93	1,476
0,14	-1,080	0,34	-0,412	0,54	0,1	0,74	0,643	0,94	1,555
0,15	-1,036	0,35	-0,385	0,55	0,126	0,75	0,674	0,95	1,645
0,16	-0,994	0,36	-0,358	0,56	0,151	0,76	0,706	0,96	1,751
0,17	-0,954	0,37	-0,332	0,57	0,176	0,77	0,739	0,97	1,881
0,18	-0,915	0,38	-0,305	0,58	0,202	0,78	0,772	0,98	2,054
0,19	-0,878	0,39	-0,279	0,59	0,228	0,79	0,806	0,99	2,326

0,991	0,992	0,993	0,994	0,995	0,996	0,997	0,998	0,999
2,366	2,409	2,257	2,512	2,576	2,652	2,748	2,878	3,090

Таблица 7

Функция распределения стандартного нормального закона

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$$

Для $x < 0$ следует пользоваться формулой: $F(-x) = 1 - F(x)$
(например, $F(-2) = 1 - F(2) = 1 - 0,977 = 0,023$)

x	F(x)	x	F(x)	x	F(x)	x	F(x)
0,0	0,500	1	0,841	2,0	0,977	3,0	0,998
0,1	0,540	1,1	0,864	2,1	0,982	3,1	0,999
0,2	0,579	1,2	0,885	2,2	0,986	3,2	0,9993
0,3	0,618	1,3	0,903	2,3	0,989	3,3	0,9995
0,4	0,655	1,4	0,919	2,4	0,992	3,4	0,9997
0,5	0,691	1,5	0,933	2,5	0,994	3,5	0,9998
0,6	0,726	1,6	0,945	2,6	0,995	3,6	0,9998
0,7	0,758	1,7	0,955	2,7	0,996	3,7	0,9999
0,8	0,788	1,8	0,964	2,8	0,997	3,8	0,9999
0,9	0,816	1,9	0,971	2,9	0,998	3,9	0,9999

Таблица 8

Значения k_β нормального распределения

β	$\alpha = 1 - \beta$	$\tilde{\alpha} = \alpha / 2$	k_β
0,9	0,1	0,05	1,65
0,95	0,05	0,025	1,96
0,98	0,02	0,01	2,3
0,99	0,01	0,005	2,58
0,9975	0,0025	0,00125	3,02

Принятые обозначения: α – ошибка, β – уровень доверия. В таблице заданы

k_β – решения уравнения $\beta = \frac{1}{\sqrt{2\pi}} \int_{-k_\beta}^{+k_\beta} e^{-\frac{x^2}{2}} dx$ (значения β в первом столбце, соот-

ветствующая ошибка $\alpha = 1 - \beta$ во втором столбце, так что если задана ошибка α , то надо искать ее во втором столбце). Если нужно строить одностороннюю об-

ласть, т.е. надо решать уравнение: $\tilde{\alpha} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-k_{\tilde{\alpha}}} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{+k_{\tilde{\alpha}}}^{\infty} e^{-\frac{x^2}{2}} dx$, то

$\tilde{\alpha} = 1 - \beta$ надо искать в третьем столбце.

Таблица 9

Значения $t_{n,\beta}$ распределения Стьюдента

Число степеней свободы	Уровень значимости $\alpha = 1-\beta$ (двусторонняя критическая область). Заданы α/β				
	n	0,10/0,90	0,05/0,95	0,02/0,98	0,01/0,99
1	6,31	12,7	31,82	63,7	
2	2,92	4,30	6,97	9,92	
3	2,35	3,18	4,54	5,84	
4	2,13	2,78	3,75	4,6	
5	2,01	2,57	3,37	4,03	
6	1,94	2,45	3,14	3,71	
7	1,89	2,36	3,00	3,5	
8	1,86	2,31	2,9	3,36	
9	1,83	2,26	2,82	3,25	
10	1,81	2,23	2,76	3,17	
11	1,80	2,220	2,72	3,11	
12	1,78	2,18	2,68	3,05	
13	1,77	2,16	2,65	3,01	
14	1,76	2,14	2,62	2,98	
15	1,75	2,13	2,60	2,95	
16	1,75	2,12	2,58	2,92	
17	1,74	2,11	2,57	2,90	
18	1,73	2,10	2,55	2,88	
19	1,73	2,09	2,54	2,86	
20	1,73	2,09	2,53	2,85	
30	1,70	2,04	2,46	2,75	
40	1,68	2,02	2,42	2,70	
60	1,67	2,00	2,39	2,66	
120	1,66	1,98	2,36	2,62	
∞	1,64	1,96	2,33	2,58	
	0,05/0,95	0,025/0,975	0,01/0,99	0,005/0,995	
Уровень значимости $\alpha = 1-\beta$ (односторонняя критическая область)					

Принятые обозначения: α – уровень значимости (ошибка), β – уровень доверия.

В таблице заданы $t_{n,\beta}$ – решения уравнения $\beta = \int_{-t_{n,\beta}}^{+t_{n,\beta}} f_n(x) dx$ (значения β в верхней

строке). Если нужно решать уравнение $\alpha = \int_{-\infty}^{-t_{n,\alpha}} f_n(x) dx = \int_{+t_{n,\alpha}}^{\infty} f_n(x) dx$, то α надо

искать в нижней строке.

Таблица 10

Критические значения λ_α критерия Колмогорова

α	0,20	0,10	0,05	0,025	0,01	0,005	0,001
λ_α	1,07	1,22	1,36	1,48	1,63	1,73	1,95

Таблица 11

**Последовательность случайных чисел,
распределенных равномерно на отрезке [0,100]**

10	09	73	25	33	76	52	01	35	86	34	67	35	48	76	80	95	90	91	17
37	54	20	48	05	64	89	47	42	96	24	80	52	40	37	20	63	61	04	02
08	42	26	89	53	19	64	50	93	03	23	20	90	25	60	15	95	33	47	64
99	01	90	25	29	09	37	67	07	15	38	31	13	11	65	88	67	67	43	97
12	80	79	99	70	80	15	73	61	47	64	03	23	66	53	98	95	11	68	77
66	06	57	47	17	34	07	27	68	50	36	69	73	61	70	65	81	33	98	85
31	06	01	08	05	45	57	18	24	06	35	30	34	26	14	86	79	90	74	39
85	26	97	76	02	02	05	16	56	92	68	66	57	48	18	73	05	38	52	47
22	15	67	16	01	76	72	52	73	62	79	88	03	40	47	40	99	58	39	51
05	94	66	77	42	77	53	12	97	87	01	95	47	73	83	68	41	90	12	26

Таблица 12

**Последовательность случайных чисел,
имеющих распределение N(0,1)**

0,414	0,011	0,666	-1,132	-0,410	-1,077	1,484	-0,340	0,789	-0,494	0,364
-1,237	-0,044	-0,111	-0,210	0,931	0,616	-0,377	-0,433	1,048	-0,037	0,759
0,609	-2,043	-2,290	0,404	-0,543	0,486	0,869	0,347	2,816	-0,464	-0,632
-1,614	0,372	-0,074	-0,916	1,314	-0,038	0,673	0,563	-0,107	0,131	-1,808
0,284	0,458	1,307	-1,625	-0,629	-0,504	-0,0056	-0,131	0,048	1,879	-1,016
0,360	-0,119	2,331	1,672	-1,053	0,840	0,246	-0,237	-1,312	1,603	-0,952
-0,566	1,600	0,465	1,951	0,110	0,251	0,116	-0,957	-0,190	1,479	-0,986
1,249	1,934	0,070	-1,358	-1,246	-0,959	-1,297	-0,722	0,925	0,783	-0,402
0,619	1,826	1,272	-0,945	0,494	0,050	-1,696	1,876	0,063	0,132	0,682
0,544	-0,417	-0,666	-0,104	-0,253	-2,543	-1,133	1,987	0,668	0,360	1,927
1,183	1,211	1,765	0,035	-0,359	0,193	-1,023	-0,222	-0,616	-0,060	-1,319
-0,785	-0,430	-0,298	0,248	-0,088	-1,379	0,295	-0,115	-0,621	-0,618	0,209
0,979	0,906	-0,096	-1,376	1,047	-0,872	-2,200	-1,384	1,425	-0,812	0,748
-1,095										

Изд. лиц. ЛР № 071765 от 07.12.1998
Казахстанско-Российский университет
Объем 10 п.л. Тираж 922

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000